



**The application and development of methods to
combine and infer information from genetic
epidemiological studies of cardiovascular and other
complex traits**

Nicholas Gareth Daniel Masca

Department of Health Sciences

May 2011

Thesis submitted for the degree of Doctor of Philosophy at
the University of Leicester

Abstract

The application and development of methods to combine and infer information from genetic epidemiological studies of cardiovascular and other complex traits.

Nicholas Gareth Daniel Masca

This thesis investigates methods to combine and infer information from genetic epidemiological studies. Three issues are explored, each in a distinct and self-contained chapter.

Chapter 1 investigates how best to incorporate treatment information in genetic analyses of blood pressure. Different approaches to adjusting for treatment are compared in a number of simulated scenarios, and the approaches that utilise all the observed data are generally shown to perform best. One particular condition, however, causes these approaches to suffer bias. This is where a genetic variant (or some other factor) interacts with treatment. This chapter therefore urges caution in the interpretation of results from these studies, and suggests some possible approaches to identifying existing interactions with treatment.

Chapter 2 concerns participant privacy in genome-wide association studies (GWAS). Recent methods claim to be able to infer whether an individual participated in a study, using only aggregate statistics from the study such as allele frequencies. In the past, these statistics have been freely published online. This chapter explores the full implications of these methods, by investigating their true capabilities and limitations. In addition, some modifications are proposed to one particular method, to demonstrate how it can be adapted for use in practice. This work finds that participant identification is possible in ideal conditions, but common characteristics of real studies may prevent any reliable application of these methods in practice.

Chapter 3 proposes a new approach to synthesising data between studies. This approach – named “DataSHIELD” – guarantees identical results to an *individual-level* meta-analysis, while offering greater flexibility than the *study-level* meta-analysis. DataSHIELD also potentially circumvents some of the laws that restrict data use, because it does not involve sharing any individual-level data between studies. This chapter outlines the principles underpinning DataSHIELD, and demonstrates its use in a simulated data example.

Acknowledgements

I would like to thank my supervisors, Nuala, Paul, and Martin, for all their help and support over the last three or so years. I am extremely grateful for the guidance you provided, and for all the knowledge and enthusiasm you all shared with me.

I would also like to thank the other PhD students and staff in the Health Sciences department for offering encouragement and advice whenever needed, and for creating a stimulating environment to work in.

I am very grateful to the British Heart Foundation for providing me with the funding to undertake to this project, which has ultimately given me this opportunity.

Finally, I would like to thank my parents for their continuous encouragement and support, and Katie, for always being there, and putting up with me!

Contents

Preface.....	1
Chapter 1.	
Correcting for the Use of Antihypertensive Treatment in Genetic Analyses of Blood Pressure.....	4
1.1.Introduction	4
1.1.1 Background	4
1.1.2 Approaches to correct for modified BPs	7
(a) No Adjustment	8
(b) Exclude	9
(c) Treatment as a Binary Covariate.....	10
(d) Binary Trait.....	11
(e) Fixed Treatment Effect.....	13
(f) Fixed Substitution.....	14
(g) Random Substitution.....	15
(h) Median Method	16
(i) Non-Parametric Adjustment	17
(j) Censored Normal Regression	19
1.1.3 Alternative Approaches	20
1.1.4 Comparison of the Approaches	22
1.1.5 Existing Work & Proposed Extensions	23

1.2.Simulation Studies: Non-differential intervention.....	27
1.2.1 General Simulation Study.....	28
1.2.1.1 Simulation Method	28
1.2.1.2 Results	33
1.2.2 Scenario 1: Unobserved Covariate.....	38
1.2.2.1 Simulation Method	38
1.2.2.2 Results	39
1.2.3 Scenario 2: Treated Normotensives	42
1.2.3.1 Simulation Method	42
1.2.3.2 Results	44
1.2.4 Scenario 3: Combination Therapy	47
1.2.4.1 Simulation Method	47
1.2.4.2 Results	49
1.2.5 Scenario 4: Proportional Treatment Effect.....	51
1.2.5.1 Simulation Method	52
1.2.5.2 Results	54
1.3. Simulation Studies: Differential intervention.....	60
1.3.1 Scenario 5: Pharmacogenetic Interaction.....	61
1.3.1.1 Simulation Method	61
1.3.1.2 Results	66

1.3.2 Scenario 6: Pharmacogenetic Interaction with One Class of Treatment.....	73
1.3.2.1 Simulation Method	73
1.3.2.2 Results	76
1.3.3 Scenario 7: Differential Probability for Receiving Treatment	79
1.3.3.1 Simulation Method	81
1.3.3.2 Results	83
1.4. Discussion.....	86
1.4.1 Summary and Explanation of the Results.....	86
1.4.2 Practical Recommendations.....	90
1.4.3 Implications	93
1.4.4 Applicability of the Findings	94
1.4.5 Conclusions.....	96
Chapter 2.	
Participant Identifiability in GWAS.....	98
2.1. Introduction	97
2.2. The Homer Method	99
2.2.1 Outline of the Method	99
2.2.2 Practical Issues	102
2.2.2.1 Reference Population.....	102
2.2.2.2 Composite Hypotheses	104

2.2.3	Different Applications.....	106
2.2.4	Assumptions.....	114
2.2.5	Other Important Characteristics.....	116
2.3.	Response to the Homer Method	118
2.3.1	Previous Findings.....	119
2.3.2	Extensions of the Homer method	121
2.3.2.1	Jacobs <i>et al.</i>	121
2.3.2.2	Visscher <i>et al.</i>	124
2.3.2.3	Sampson & Zhao	125
2.3.2.4	Clayton	127
2.3.2.5	Sankararaman <i>et al.</i>	129
2.4.	Testing the Original Homer Method	131
2.4.1	Simulation Method.....	131
2.4.2	Scenario 1: Number of SNPs	134
2.4.3	Scenario 2: Reference Group and Mixture Size	136
2.4.4	Discussion of the Results	139
2.5.	Visscher <i>et al.</i> Linear Regression.....	142
2.5.1	Overview	143
2.5.2	Assumptions and Practical Implications	147
2.6.	Testing the Visscher <i>et al.</i> Method: Simulation Studies	149
2.6.1	Scenario 1: Random Sampling.....	150

2.6.2	Scenario 2: Sampling by Disease Status.....	153
2.7.	Real Data Illustrations & Extensions	159
2.7.1	1958 Birth Cohort	160
2.7.2	Common Ancestry & No LD	161
2.7.3	Modelling the variance.....	170
2.7.3.1	Logistic Regression	170
2.7.3.2	GEE Independence Model	173
2.7.3.3	Testing the Models.....	177
2.7.4	Adjusting for LD	183
2.7.4.1	GEE Model with AR-1 Correlation Structure	184
2.7.4.2	Testing the GEE AR-1 Model	185
2.7.4.3	Influence of Cluster Size	189
2.8.	Ancestry	196
2.8.1	Previous Findings and Preliminary Results	197
2.8.2	Simulation Study.....	199
2.8.3	Comparing UK Regions.....	206
2.8.4	Comparing Different UK Cohorts.....	209
2.9.	What can be published?.....	212
2.9.1	Limiting the number of SNPs.....	212
2.9.2	Other types of summary information.....	213
2.9.3	Sign Test	215

2.9.4	Implications	218
2.10.	Discussion.....	219
2.10.1	Heteroscedasticity	220
2.10.2	The implications of linkage disequilibrium (LD)	221
2.10.3	Ancestry	223
2.10.4	Forensic use of the tests	224
2.10.5	Conclusions.....	224
Chapter 3.		
	A new approach to data synthesis: DataSHIELD	227
3.1.	Introduction	227
3.1.1	Existing approaches to data synthesis	228
3.1.1.1	Study Level Meta Analysis	229
3.1.1.2	Individual Level Meta Analysis	232
3.1.1.3	Comparison of the approaches	233
3.1.2	Ethico-legal issues surrounding the sharing of data	234
3.1.3	What is needed?.....	237
3.2.	DataSHIELD.....	238
3.2.1	What is DataSHIELD?	239
3.2.2	Deriving descriptive statistics in DataSHIELD	243
3.2.3	Fitting a linear model in DataSHIELD	244
3.2.4	Fitting a GLM in DataSHIELD.....	246

3.2.4.1 The IRLS Algorithm.....	248
3.2.4.2 Applying the IRLS algorithm to horizontally-partitioned data...	251
3.2.5 Key requirements	253
3.3. Simulation Studies	256
3.3.1 Scenario 1: Normally distributed data.....	258
3.3.1.1 Simulation method	258
3.3.1.2 Approach to Analysis.....	259
3.3.1.3 Results	259
3.3.2 Scenario 2: Binary data	260
3.3.2.1 Simulation Method	260
3.3.2.2 Approach to Analysis.....	263
3.3.2.3 Results	265
3.4. Discussion.....	266
3.4.1 Further ethico-legal issues	266
3.4.2 The IT system.....	267
3.4.3 Further developments.....	269
3.4.4 Conclusions.....	269
Conclusions and Further Work.....	271
Chapter 1.....	272
Chapter 2.....	273
Chapter 3.....	275

Final Conclusions	277
Appendix A	279
Appendix B.....	301
B.1. NIH Background Fact Sheet on GWAS Policy Update	301
B.2... Breakdown and Proof of the Visscher <i>et al.</i> Linear Regression Approach	303
Appendix C.....	313
C.1. Scenario 1	313
C.1.1 R code for simulating the data.....	313
C.1.2 R Code & Output for Analysis 1 (ILMA):.....	314
C.1.3 R code & Output for Analysis 2 (DataSHIELD analysis):.....	314
C.2. Scenario 2	319
C.2.1 R code for simulating the data.....	319
C.2.2 R Code & Output for Analysis 1 (ILMA):.....	322
C.2.3 R Code to perform Analysis 2 (DataSHIELD analysis):.....	323
C.2.4 Output from Analysis 2:	341
C.2.5 Final Results for Analysis 2:	350
Bibliography	351
Appendix.....	367

List of Tables

Table 1: Common assumptions used for approaches to correct for modified phenotypes.....	8
Table 2: Simulation Properties for the General Simulation Study.....	30
Table 3: Summary of the analysis models with parameter values used in the simulated studies.	32
Table 4: Descriptive Statistics for 1,000 datasets of the General Simulation Study.....	33
Table 5: Simulation Properties for Scenario 1.....	39
Table 6: Simulation properties for Scenario 2.	43
Table 7: Descriptive statistics for Scenario 2.....	44
Table 8: Simulation properties for Scenario 3.	48
Table 9: Descriptive statistics for Scenario 3.....	49
Table 10: Simulation Properties for Scenario 4.....	53
Table 11: Descriptive statistics for Scenario 4.....	54
Table 12: Simulation properties for Scenario 5.	65
Table 13: Descriptive statistics for Scenario 5a.....	66
Table 14: Simulation Properties for Scenario 6.....	75
Table 15: Descriptive statistics for Scenario 6a.....	76
Table 16: Simulation Properties for Scenario 7.....	82
Table 17: Descriptive Statistics for Scenario 7.....	83

Table 18: Results for 100 runs of Scenario One..	152
Table 19: Simulation characteristics for each simulated case-control GWAS.	156
Table 20: Results for Scenario Two..	157
Table 21: Results for SNP spacing of 20, using individuals only from southern UK regions.	162
Table 22: Results for the Visscher et al. linear regression approach applied to the 1958BC data with SNP spacing of 20, 33 and 100..	167
Table 23: Results for analyses of the 1958BC using SNP spacing of 100 and sampling only individuals with southern UK ancestry.....	179
Table 24: Results for analyses of the 1958BC using SNP spacing of 20 and sampling only individuals with southern UK ancestry.....	182
Table 25: Analyses of the 1958 Birth Cohort data with SNP spacing of 20....	188
Table 26: Results for analyses of 1958BC data – chromosome 14.....	192
Table 27: Results for the simulation study investigating the effects of population divergence.....	203
Table 28: Comparison of regions in the 1958 Birth Cohort with SNP spacing of 20	208
Table 29: Comparison of Different UK Cohorts.....	211
Table 30: Results for an ILMA analysis (Analysis 1) and a DataSHIELD SLMA analysis (Analysis 2) of the simulated SBP data.	260
Table 31: Numbers of cases and controls in the six simulated studies.	263

Table 32: Results for an ILMA analysis (Analysis 1) and a DataSHIELD analysis (Analysis 2) of the six MI case-control studies.	265
Table 33: Results for 1,000 runs of the General Simulation Study.....	280
Table 34: Results for 1,000 runs of Scenario 1.	281
Table 35: Results for 1,000 runs of Scenario 2.	282
Table 36: Results for 1,000 runs of Scenario 3.	283
Table 37: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 5%.	284
Table 38: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 10%.	285
Table 39: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 15%.	286
Table 40: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 20%.	287
Table 41: Results for 1,000 runs of Scenario 5a when $\beta_3 = 2$	288
Table 42: Results for 1,000 runs of Scenario 5a when $\beta_3 = -2$	289
Table 43: Results for 1,000 runs of Scenario 5a when $\beta_3 = 0$	290
Table 44: Results for 1,000 runs of Scenario 5b when $\beta_3 = 2$	291
Table 45: Results for 1,000 runs of Scenario 5b, when $\beta_3 = -2$	292
Table 46: Results for 1,000 runs of Scenario 5b, when $\beta_3 = 0$	293
Table 47: Results for 1,000 runs of Scenario 6a, when $\beta_3 = 2$	294
Table 48: Results for 1,000 runs of Scenario 6a, when $\beta_3 = -2$	295

Table 49: Results for 1,000 runs of Scenario 6a, when $\beta_3 = 0$	296
Table 50: Results for 1,000 runs of Scenario 6b, when $\beta_3 = 2$	297
Table 51: Results for 1,000 runs of Scenario 6b, when $\beta_3 = -2$	298
Table 52: Results for 1,000 runs of Scenario 6b, when $\beta_3 = 0$	299
Table 53: Results for 1,000 runs of Scenario 7, when the probability of treatment is differential by exposure to diabetes.....	300

List of Figures

Figure 1: Graphical representation of results for the General Simulation Study.	35
Figure 2: Graphical representation of the results for Scenario 1..	41
Figure 3: Graphical representation of results for Scenario 2..	45
Figure 4: Graphical representation of results for Scenario 3..	50
Figure 5: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 5%..	57
Figure 6: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 10%..	58
Figure 7: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 15%..	59
Figure 8: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 20%..	60
Figure 9: Graphical representation of the results for Scenario 5a..	68
Figure 10: Graphical representation of the results for Scenario 5b..	69
Figure 11: Graphical representation of the results for Scenario 6a..	78
Figure 12: Graphical representation of the results for Scenario 6b..	79
Figure 13: Graphical representation of the results for Scenario 7 ..	84
Figure 14: Illustration showing the relationship between the individuals in the mixture, the reference group, and the test individuals who are in neither group	132

Figure 15: ROC curve showing the sensitivity and 1-specificity of the Homer method when various numbers of SNPs (s) are used.....	135
Figure 16: Histograms of the test statistic under the alternative and null hypotheses.....	137
Figure 17: Histograms of the test statistic under the alternative and null hypotheses.....	139
Figure 18: Histogram of the z-statistic under the null hypothesis..	164
Figure 19: Histogram of allele frequencies (AFs) for every 20 th SNP on chromosomes 1-10.	169
Figure 20: Schematic representation of the ILMA (a) and SLMA (b) approaches.	233
Figure 21: Illustration of the DataSHIELD IT infrastructure (taken from (Wolfson <i>et al.</i> , 2010)).....	241
Figure 22: Flowchart representing the processes involved in fitting a GLM in DataSHIELD.....	253

Preface

This thesis follows a general theme concerning the development and application of statistical methodology for combining and inferring information from genetic epidemiological studies of cardiovascular and other complex traits. Three chapters approach the issue from different angles and address the theme in different ways. As such, each chapter is self-contained, and includes its own methods and results, as well as specific introductory and discussion sections.

Chapter 1 explores an issue that is particularly apparent in genetic epidemiological analyses of blood pressure. This chapter focuses on how to adjust an analysis when some observations are distorted by the use of treatment, and is an example of incorporating treatment information. Different approaches to handling these observations – which may be considered to be “right censored” – are compared, with the aim of recommending the most appropriate methods to use in different settings. This chapter forms the basis of a paper published in *Statistics in Medicine* (Masca *et al.*, 2011).

In Chapter 2 the issue of inferring information is addressed in the context of the publication and use of results from genome-wide association studies. The release of results is a vital aspect of the research process; for example, it allows

for the verification of findings and it can inform further work. However, recent assertions threaten to limit what can be published safely from these studies, in light of the ethico-legal requirements to guarantee the protection of participant confidentiality. For instance, methods have been proposed that claim to be able to test probabilistically whether or not a given individual of interest participated in a particular study using only aggregate statistics from studies (such as allele frequencies). These findings have had major implications on the data sharing practices in the field, but a number of issues have remained unclear. As such, Chapter 2 explores the science behind these methods with the aim of clarifying their true capabilities and limitations. A paper based on the findings in this chapter has been submitted to the *International Journal of Epidemiology* and is currently in press.

Chapter 3 concerns the combining of data and results across studies. Existing approaches to synthesising data, such as study level and individual level meta-analysis, either lack flexibility or can infringe upon the ethico-legal stipulations that restrict data use. A need for a more flexible approach that avoids contravening these data sharing laws therefore exists. Chapter 3 outlines such an approach, which has been published in the *International Journal of Epidemiology* (Wolfson *et al.*, 2010). I am a co-author on that paper. This approach involves the use of a dedicated IT infrastructure and specialised statistical algorithms to permit the pooling of results across studies without the need to share any *individual-level* data. As such, it adheres to the strict data privacy standards required in the field, and allows improved flexibility to specify and execute analyses from a single research hub. I contributed to this paper primarily in terms of the development of the mathematical models underpinning

the approach. I also drafted the supplementary materials for the paper, which detail the mathematical algorithms and the programming code.

The thesis concludes with a general discussion highlighting areas for further work and possible extensions. The appendices contain supplementary material, where appropriate, including results tables, computer code, and mathematical proofs. The Appendix is split into three sections: A, B, and C, which contain relevant information for chapters 1, 2 and 3 respectively.

Chapter 1.

Correcting for the Use of Antihypertensive Treatment in Genetic Analyses of Blood Pressure

1.1. Introduction

This chapter investigates a problem that occurs with analyses that focus on the aetiology of certain complex traits. The problem arises in observational studies where a number of the participants use a form of treatment that directly impacts upon the observed outcome of interest. Studies into cardiovascular traits for which treatment is widely prescribed - such as blood pressure and high or low-density lipoprotein – are therefore particularly affected by this issue. This chapter primarily investigates the issue in the context of studies of blood pressure (BP); however, it is important to note that the key findings generalise to the analysis of other traits that are mitigated by treatment.

1.1.1 Background

Hypertension (high BP) is a common condition estimated to affect over 25% of adults worldwide (Kearney *et al.*, 2005). Although hypertension itself is asymptomatic, it is a major contributor to the risk of cardiovascular disease,

which accounts for up to 30% of all deaths (Murray *et al.*, 1997; Chobanian *et al.*, 2003). Even changes within the normal range BP are associated with risk of stroke and coronary heart disease (Lewington *et al.*, 2002). BP in its own right is therefore of major importance to public health.

Lifestyle factors such as dietary salt intake, physical activity, smoking, and body-mass index (BMI) are all known to influence BP (Beilin, 1997; Pickering, 1997), but BP also has a substantial heritable component (Havlik *et al.*, 1979; Levy *et al.*, 2000). Identification of the genetic determinants of BP can offer insights into the biological pathways underpinning BP regulation (Lifton *et al.*, 2001), and, indeed, this has been a key aim of recent genetic association studies of BP.

Paramount to the success of a genetic association study is a sufficient statistical power to detect the generally modest effects of common genetic variants (Wong *et al.*, 2003; Burton *et al.*, 2009). In genome-wide association studies, hundreds of thousands or even millions of genetic variants are tested for association with the phenotype of interest, and an allowance for multiple testing must be made. As such, the threshold for genome-wide significance is usually currently defined as $p < 5 \times 10^{-8}$ (McCarthy *et al.*, 2008). A sufficiently large sample size is crucial to the provision of an adequate power to detect associations for BP at this threshold. Recent breakthroughs in genome-wide association studies of BP have been achieved using large sample sizes (Levy *et al.*, 2009; Newton-Cheh *et al.*, 2009). For instance, the Global BPgen Consortium (Newton-Cheh *et al.*, 2009) meta-analysed 17 cohorts consisting of a total of 34 433 participants, and the CHARGE Consortium (Levy *et al.*, 2009) meta-analysed five cohorts

consisting of 29 136 participants. The SNPs highlighted in these studies had reported effect sizes between approximately 0.5 and 1 mmHg per copy of the minor allele for systolic BP, and approximately 0.35 to 0.5 mmHg per copy of the minor allele for diastolic BP (typically about 1/40th to 1/15th of a standard deviation).

Besides sample size, there are several other factors that may limit the statistical power of genetic association studies of BP. For instance, it can be difficult to gain a reliable measure of an individual's BP (Wong *et al.*, 2003) because BP varies in different situations and at different time points throughout the day. Other measurement difficulties, such as an alerting (or "white-coat") response, and observer bias (including "digit preference", which entails rounding BP readings up or down) can also influence recordings of BP (Wilcox, 1961; Petrie *et al.*, 1986). Most importantly, investigations into the aetiology of BP are affected by the use of antihypertensive treatments by study participants. Since hypertension is highly prevalent within western countries, drugs to lower BP – antihypertensives – are widely prescribed. Population-based cohort studies therefore sometimes have up to a quarter of participants on antihypertensive treatment (or even more in studies of older populations) (Levy *et al.*, 2009; Newton-Cheh *et al.*, 2009). For these treated participants, any BP measurements provided in a study will reflect "modified BP" values, as opposed to the "underlying BP" values that exist, in principle, in the absence of treatment.

It has been shown that a failure to adequately correct analyses for the inclusion of modified BPs can distort the results (White *et al.*, 1994; Cook, 1997; White *et al.*, 2003; Tobin *et al.*, 2005; McClelland *et al.*, 2008). Because

antihypertensive treatments lower BP, *modified BPs* will be lower than the unobserved, *underlying BPs*, and the results of an analysis may thus be misleading if no account is made for antihypertensive use. A number of approaches have therefore been proposed to adjust for such bias. The following section introduces these approaches.

1.1.2 Approaches to correct for modified BPs

Before describing the approaches to correct for modified BPs, I first introduce some notation, which shall be used throughout the chapter.

For the i^{th} subject ($i = 1, \dots, n$), Y_i is the *observed* systolic blood pressure (SBP), and Z_i is a latent variable representing the *underlying* SBP. For subjects who use antihypertensive treatment, Z_i cannot be observed, so it is estimated or imputed following an algorithm defined individually by each approach. Imputed values of Z_i are denoted by the variable X_i . The indicator variable $TREAT_i$ is used to denote whether the i^{th} subject receives treatment ($TREAT_i = 1$ if treated; $TREAT_i = 0$ otherwise). Note that for all approaches and in all situations, if $TREAT_i = 0$ then $X_i = Y_i = Z_i$.

Each of the approaches, unless stated otherwise, fits a linear regression model to X_i . I consider the same model for each approach, consisting of an age effect; a sex effect; and a genetic factor. The approaches therefore fit a model to the imputed values of the form:

Equation 1

$$X_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 g_i + \varepsilon_i,$$

where AGE_i is a continuous covariate measured in years, SEX_i is a binary covariate ($SEX_i = 1$ denotes a male, $SEX_i = 0$ denotes a female), and g_i denotes the genotype for a diallelic locus ($g_i = 0, 1$ or 2 copies of the minor allele). The term ε_i represents random error, and $\varepsilon_i \sim N(0, \sigma^2)$. Note that an additive genetic effect is fitted in the above model, where two copies of the minor allele produce twice the effect of one copy.

Each of the approaches to correct for modified BPs relies on a set of assumptions, and there are common classes of assumption shared by the different approaches. Table 1 below therefore lists the main assumption classes. For convenience, I later refer back to these assumptions, where applicable, using the class numbers provided. Note that all approaches that fit a linear regression model rely on Assumption (i), and I therefore shall not further allude to any reliance on this assumption further.

Class	Assumption
(i)	Assumptions linked to the type of model fitted. For instance, any approach that fits a linear regression assumes that the error terms are independent and follow a normal distribution with constant variance.
(ii)	Assumption regarding the size/nature of the treatment effect.
(iii)	Assumption regarding the distribution of the underlying phenotype.
(iv)	Assumption regarding whether or not the modified phenotypes are informative of underlying patterns/phenotypes.

Table 1: Common assumptions used for approaches to correct for modified phenotypes.

(a) No Adjustment

A common approach to analysis is simply to ignore the problem altogether. Hence, no correction for the use of treatment is implemented, and all the

observed BPs are analysed in a conventional way. A model in the form of Equation 1 is therefore fitted, but where the estimates of the underlying phenotype, X_i , are simply equal to the observed phenotypes, Y_i :

Equation 2

$$X_i \equiv Y_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 g_i + \varepsilon_i$$

The *No Adjustment* approach assumes that modified BPs (i.e. the observed BPs for individuals on treatment) are informative of underlying patterns [Assumption (iv)]. This is an implicit assumption, because the approach actually ignores the problem completely.

(b) Exclude

The *Exclude* approach assumes that modified BPs are uninformative of the underlying BPs [Assumption (iv)], and omits any treated individuals from the analysis. It therefore fits a model in the form of Equation 1 only to the remaining subjects, i.e. for which $treat_i = 0$.

The *Exclude* approach is commonly used in practice (e.g. (Hsueh *et al.*, 2000; Rice *et al.*, 2000; Brand *et al.*, 2003)). It is inefficient, however, in that it typically suffers an inevitable loss of statistical power as a consequence of disregarding a possibly sizeable proportion of the data.

The *Exclude* approach can alternatively be performed by omitting any individuals on treatment at the recruitment stage of a study. In this situation, a target sample size is acquired and, hence, any power loss due to excluding data from treated subjects is avoided. Even this strategy remains

unsatisfactory, however. Such a recruitment strategy imposes a selection bias upon a study, in which the sample becomes biased towards individuals with lower underlying BPs. Given that individuals on treatment tend to be those with high underlying BP – and that these are potentially “interesting” subjects – this alternative strategy can clearly be seen to be flawed.

(c) Treatment as a Binary Covariate

Another approach commonly used in practice is to adjust for treatment by modelling it as a binary covariate (e.g. (Yang *et al.*, 2007; Vora *et al.*, 2008)). This requires fitting a model in the form of Equation 2, with the additional term $TREAT_i$:

Equation 3

$$X_i \equiv Y_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 TREAT_i + \beta_4 g_i + \varepsilon_i$$

Treatment as a Binary Covariate assumes that modified phenotypes are informative of underlying patterns as long as treatment is adjusted for in the model [Assumption (iv)]. Although this approach, at first, seems reasonable, as we shall see, on closer inspection it is actually flawed.

Consider that the aim of a study of BP is to investigate the determinants of BP. Where an analysis adjusts for treatment by modelling it as a covariate, in effect, it explains away the differences in BP between subjects who use treatment and subjects who do not by attributing these differences to an apparent “treatment effect”. This is an inappropriate strategy because the differences in BP between subjects are not caused by treatment at all (treatment actually *reduces* the differences between those with high and low BP). Estimating the effects of

the factors that *truly* cause these differences should be the primary focus of the analysis. Hence, attributing these differences to treatment explains away variation in the data and could mask any true effects.

The problem with this analysis can further be explained by considering the role of treatment in this scenario. Where conventional covariates are defined as “possible predictors” of the outcome of interest (Last, 2001), the use of treatment here not only *predicts* BP – but also is a *consequence* of having high BP. Therefore, treatment is not a conventional covariate here, and should not be adjusted for in a regression in the usual way.

(d) Binary Trait

The *Binary Trait* approach classifies subjects either as *affected* or *unaffected* with regard to the condition of interest. For a study of BP, a subject would be labelled *hypertensive* if he/she uses antihypertensive treatment or, following criteria outlined in the Seventh Report of the Joint National Committee (JNC VII) (Chobanian *et al.*, 2003), for example, if he/she has SBP/DBP equal to or above 140/90 mmHg. All other subjects would be labelled “normotensive”.

The binary outcome, *hypertension_i* (= 1 if hypertensive; = 0 if normotensive) is fitted in a logistic regression model of the form:

Equation 4

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 g_i + \epsilon_i$$

where p_i is the probability that the i 'th patient is hypertensive, and all other notation is as before. Model parameters are estimated in the form of log-odds ratios.

Binary Trait assumes that any subject on treatment has the condition of interest [Assumption (iii)]. Hence, for BP, this approach assumes that anyone who uses antihypertensive medication has hypertension. Any subjects on treatment that do not fit the conventional criteria for treatment can therefore pose problems. For example, individuals with diabetes are more susceptible to the risks associated with high BP and, hence, are often prescribed antihypertensives at a lower threshold than that for non-diabetics. Similarly, antihypertensive medications are sometimes prescribed for conditions such as coronary heart disease (CHD), heart failure and migraine, rather than to treat hypertension. Clinically normotensive individuals, thus, may also use antihypertensives, and the *Binary Trait* approach, in effect, misclassifies these individuals. This could be a source of bias. Note that any possible misclassifications depend on the definition of hypertension. Although the above definition is common, other definitions could also be used instead.

A further limitation of the *Binary Trait* approach is that, typically, dichotomising a continuous outcome leads to a loss of statistical power (Altman *et al.*, 2006). *Binary Trait* is therefore likely to make inefficient use of the quantitative BP measurements collected.

(e) Fixed Treatment Effect

This approach is outlined by Cui et al. (Cui *et al.*, 2002; Cui *et al.*, 2003). They suggest adding a fixed constant, c , to all modified BPs to adjust for the negative effect of treatment. Cui et al. recommend that the choice of c is based on an average effect for the appropriate class of treatment, as portrayed in the medical literature. For example, an average reduction in blood pressure attributed to antihypertensive medication is typically around 10 mmHg (Law *et al.*, 2003).

This approach therefore fits a model in the form of Equation 1, where:

Equation 5

$$X_i = \begin{cases} Y_i + c & \text{if } treat_i = 1 \\ Y_i & \text{if } treat_i = 0 \end{cases}$$

Fixed Treatment Effect assumes a fixed distribution for the treatment effect [Assumption (ii)], and this is a potential weakness of the approach. For instance, some patients could be treated more aggressively, with a higher dosage of a drug or with “combination therapy” (see Section 1.2.4). In these situations, the above assumption will be violated.

The *Fixed Treatment Effect* approach implicitly assumes that modified BPs are informative [Assumption (iv)]. It therefore uses the information in the original observations, and, hence, retains the original variability in the data.

(f) Fixed Substitution

Hunt et al. (Hunt *et al.*, 2002) propose an approach that substitutes all modified BPs for a fixed value, m . They recommend setting m as the minimum threshold for the clinical diagnosis of the condition of interest. According to the JNC VII (Chobanian *et al.*, 2003), a diagnosis of hypertension is where systolic blood pressure (SBP) is greater than 140 mmHg, and/or diastolic blood pressure (DBP) is greater than 90 mmHg. Hence, for a study of BP, the recommended value for m is 140/90 for SBP/DBP respectively.

Fixed Substitution involves fitting a model in the form of Equation 1, where:

Equation 6

$$X_i = \begin{cases} m & \text{if } treat_i = 1 \\ Y_i & \text{if } treat_i = 0 \end{cases}$$

The *Fixed Substitution* approach assumes that any modified BP is uninformative of the underlying BP [Assumption (iv)]. It therefore substitutes modified BPs and, hence, removes any distortion due to treatment. This removes the variability in the original data. *Fixed Substitution* also implicitly assumes that individuals who use treatment have the condition of interest, i.e. hypertension in this case [Assumption (iii)]. As already described for (d) above, this assumption may not always hold in practice; for example, some subjects may be on antihypertensive medication for some reason other than to treat hypertension. Thus, the reliance on this assumption could be problematic under some circumstances.

(g) Random Substitution

As an alternative to (f), Hunt et al. (Hunt *et al.*, 2002) also propose replacing modified BPs with randomly generated values from a pre-specified distribution. This distribution should be centred between the typical thresholds for the diagnosis of the condition of interest, and truncated at each end. For instance, following the JNC VII (Chobanian *et al.*, 2003), stage I hypertension is defined as SBP/DBP between 140/90 mmHg and 160/100 mmHg. Hence, for BP, Hunt et al. recommend that X_i is generated randomly from a normal distribution with mean 150/95, standard deviation 5/2.5, and with truncation at 140/90 mmHg and 160/100 mmHg for SBP/DBP measures respectively.

Hence, this approach fits a model of the form in Equation 1, where:

Equation 7

$$X_i = \begin{cases} \sim N(150, 5^2) & \text{if } TREAT_i = 1 \\ Y_i & \text{if } TREAT_i = 0, \end{cases}$$

and where X_i is truncated at 140 and 160 mmHg for subjects on treatment.

Random Substitution relies on the same set of assumptions as (f). Modified BPs are assumed to be uninformative of the underlying BPs [Assumption (iv)]; and any subjects who use treatment are assumed to be hypertensive [Assumption (iii)]. The same criticisms for (d) and (f) therefore also apply here. In situations where, for example, diabetics are included within a study, some treated subjects may not necessarily be hypertensive. Assumption (iii) therefore may not be met in practice.

(h) Median Method

The *Median Method* proposed by White et al. (White et al., 1994; White et al., 2003) uses a quantile regression (Narula et al., 1999; Koenker, 2008), which models the *median* as the measure of location as opposed to the *mean* used in ordinary least squares regression. The median is less influenced by extreme observations than the mean, and is therefore generally regarded as more robust.

In contrast to ordinary least squares regression, where model residuals are assumed to be independent and normally distributed with a mean of zero and constant variance, quantile regression assumes only that the model residuals have a median of zero. Furthermore, where ordinary least squares regression estimates model parameters by minimising the sum of the squared residuals, quantile regression estimates parameter coefficients by minimising the sum of *absolute* residuals. This is a valid approach for estimating parameter coefficients because the value that minimises the sum of absolute residuals is the median.

The *Median Method* is a sign-based approach that assumes all individuals receiving treatment have an underlying BP above the median. It is claimed that where modified BPs are substituted with the constant, k (see below), then, as long as fewer than half the individuals in a study use treatment, the *Median Method* will recover the true median using quantile regression.

The *Median Method* fits a quantile regression to Equation 1, where

Equation 8

$$X_i = \begin{cases} k & \text{if } TREAT_i = 1 \\ Y_i & \text{if } TREAT_i = 0 \end{cases}$$

The estimated parameter coefficients should be insensitive to the choice of k as long as less than 50% of the individuals in a sample are treated, and that k is greater than all fitted medians. The authors recommend using a “clinically plausible value of k near the upper end of the distribution”, such as 160-200mmHg for SBP (White *et al.*, 2003). This is because they suggest estimating standard errors via bootstrap methods, which are dependent on k . Specifically, the case-resampled (or *random-x*) bootstrapping approach is recommended for estimating the standard errors, as this approach allows for any possible heterogeneity in the model residuals.

The *Median Method* assumes that modified BPs are uninformative [Assumption (iv)]. Also, as stated above, it assumes that individuals receiving treatment have an underlying BP above the median [Assumption (iii)]. This assumption could be violated in the situation described previously [see (d), (f) and (g)], where diabetics are prescribed antihypertensives at a different threshold to non-diabetics. Furthermore, because the underlying BP is unknown for subjects on treatment, in most situations Assumption (iii) cannot be validated.

(i) Non-Parametric Adjustment

Levy *et al.* (Levy *et al.*, 2000) describe a non-parametric algorithm for adjusting modified BPs for the effect of treatment. Raw residuals, r_i , are obtained by fitting the null model to the observed values, i.e.:

$$Y_i = \beta_0 + r_i.$$

The raw residuals are then sorted into descending order, and the adjusted residuals, r_k^* , calculated by applying the algorithm

Equation 9

$$r_k^* = r_k(1 - \text{treat}_k) + \text{treat}_k \left[(r_k + \sum_{j=1}^{j=k-1} r_j^*) / k \right],$$

where r_k is the k^{th} residual sorted in descending order, and treat_k is 1 if the k^{th} ordered residual relates to a treated patient, and 0 otherwise.

For untreated individuals, the adjusted residual, r_k^* , is simply equal to r_k and, hence, the algorithm shown in Equation 9 adjusts observations only for individuals who receive treatment. For treated individuals, the adjusted residual is an average of the current raw residual, r_k , together with all greater *adjusted* residuals, $\sum_{j=1}^{j=k-1} r_j^*$. The algorithm thus increases the size of the residual for any treated individual (with the exception of the individual with the greatest Y_j , if this person is treated).

Finally, an adjusted phenotype, X_k , is obtained for each individual by adding the difference between the raw and adjusted residuals, i.e.

Equation 10

$$X_k = Y_k + r_k^* - r_k.$$

A model in the form of Equation 1 is then fitted to the X_k .

Non-Parametric Adjustment assumes that modified phenotypes are informative of the underlying phenotype [Assumption (iv)], and the algorithm retains the original ordering of observations within each treatment group.

(j) Censored Normal Regression

Censored Normal Regression assumes that individuals receiving treatment are right-censored at the observed phenotype, Y_i (Tobin *et al.*, 2005). It also assumes non-informative censoring, where, conditional upon covariates, the distribution of the underlying phenotype above the censoring threshold is the same for treated and untreated participants. This allows a Tobit-type model to be fitted (Clayton *et al.*, 1993).

Based on the assumption that

$$\begin{cases} X_i \geq Y_i & \text{if } TREAT_i = 1 \\ X_i = Y_i & \text{if } TREAT_i = 0, \end{cases}$$

Tobit models use maximum likelihood to estimate the marginal effects of the model parameters (Hayashi, 2000). Hence, where the model in Equation 1 can be re-expressed as $X_i = \beta_0 + age_i\beta_1 + sex_i\beta_2 + g_i\beta_3 + \varepsilon_i = \mathbf{d}_i\boldsymbol{\beta} + \varepsilon_i$, i.e. where \mathbf{d}_i denotes the covariates (or *design* matrix) and $\varepsilon_i \sim N(0, \sigma^2)$, the contribution to the likelihood is:

$$L(y_i|d_i) = \begin{cases} \frac{1}{\sigma} \phi\left(\frac{y_i - d_i\boldsymbol{\beta}}{\sigma}\right), & \text{if } TREAT_i = 0 \text{ (Note that this is the normal p.d.f.)} \\ \Phi\left(\frac{k_i - d_i\boldsymbol{\beta}}{\sigma}\right), & \text{if } TREAT_i = 1, \text{ and where } k_i \text{ is the censoring value } Y_i. \end{cases}$$

(This is the integrated normal p.d.f. to the right of the observation, Y_i)

The log-likelihood for the i^{th} individual is therefore

Equation 11

$$\log l(y_i | d_i) = (1 - TREAT_i) \cdot \log \left[\frac{1}{\sigma} \phi \left(\frac{y_i - d_i \beta}{\sigma} \right) \right] + TREAT_i \cdot \log \left[\Phi \left(\frac{k_i - d_i \beta}{\sigma} \right) \right].$$

Censored Normal Regression relates closely to the Non-Parametric Adjustment approach detailed previously. For instance, where *Non-Parametric Adjustment* averages over the normal p.d.f. to the right of the observed value Y_i , *Censored Normal Regression* integrates over this function. The *Censored Normal Regression* approach therefore retains the assumption of normality throughout. This can be considered a distinct advantage of *Censored Normal Regression*, for the analysis of normally distributed traits such as blood pressure.

This approach assumes that modified phenotypes are informative of the underlying phenotype [Assumption (iv)]. As stated earlier, *Censored Normal Regression* also assumes non-informative censoring [Assumption (iii)]. This latter assumption is unlikely to be true in reality, because, as subjects usually only receive antihypertensive treatment if they have high BP, the distribution of underlying BP above a particular threshold will be different for treated and untreated individuals. Despite this, previous work suggests that this approach is relatively robust to this assumption.

1.1.3 Alternative Approaches

In addition to the above methods, a number of alternative approaches have been proposed that have been referred to as “Multiple Imputation” methods (Buuren *et al.*, 1999; McClelland *et al.*, 2008). The Multiple Imputation

approaches require use of additional out-of-study or pre-treatment patient records for a proportion of the subjects on treatment (Cook, 1997; Cook, 2006; McClelland *et al.*, 2008). Where available, these pre-treatment measures of BP, P_i , are assumed either to be correlated with the underlying phenotype, Z_i , (e.g. (Cook, 1997; Cook, 2006)) or to be equal to Z_i itself (e.g. (McClelland *et al.*, 2008)). Hence, for a treated subject, X_i is either a *function* of P_i , or X_i equals P_i , while for non-treated subjects, X_i is simply equal to the observed phenotype, Y_i .

Typically, pre-treatment records are only available for a proportion of the subjects in a study, and these approaches use multiple imputation to estimate P_i where absent (Rubin, 1987). For instance, the available pre-treatment measures, P_i , are fitted in a linear regression model, and covariates such as the post-treatment phenotype, medication type, dose, age, sex, and race are considered. The final model is then used to predict P_i for each patient with only post-treatment phenotypic measures available.

Although these approaches seem effective at adjusting for the use of treatment (Cook, 1997; Cook, 2006; McClelland *et al.*, 2008), a major limitation to their use is that they can only be applied in situations where some pre-treatment measures are available. Typically, pre-treatment measures are only available in longitudinal studies. Many studies of BP – such as the Global BPgen consortium (Newton-Cheh *et al.*, 2009) and the CHARGE Consortium (Levy *et al.*, 2009) – have little or no longitudinal data and, hence, the Multiple Imputation approaches are often inapplicable. The focus of this chapter is on approaches that can be applied in cross-sectional data alone. I therefore do not

consider use of the Multiple Imputation approaches any further. Note, however, that a number of the approaches described in Section 1.1.2 are basically imputation methods, because they impute BP for those measures distorted by treatment. The approaches described in Section 1.1.2 should not be confused with the “Multiple Imputation” discussed here.

1.1.4 Comparison of the Approaches

A convenient way to compare the approaches to analysis is to classify them according to how they handle any modified BPs. For instance, *No Adjustment* (a), *Exclude* (b), and *Treatment as a Binary Covariate* (c) employ either a simple correction for the use of treatment or no correction at all. As such, these approaches have been referred to as “Naïve” approaches (McClelland *et al.*, 2008).

Intuitively, each of the remaining approaches can be classified into one of two other groups. Some of the approaches assume that any modified BPs are uninformative for the underlying BPs, and they therefore substitute modified BPs for alternative values. These approaches, thus, can be described as “Substitution” approaches. Naturally, *Fixed Substitution* (f), *Random Substitution* (g), and *Median Method* (h) all fall into this class. In addition, I also classify *Binary Trait* (d) as a Substitution approach. Although (d) does not necessarily assume that modified BPs are uninformative, as with (f), (g) and (h) it substitutes modified (as well as “un-modified”) BPs for an alternative (binary) measure. Hence, even though (d) models a different trait to all other approaches, it is similar to the other approaches in this group. The Substitution approaches generally assume that any participants on antihypertensive

medication are hypertensive, and, as such, handle all modified BPs in the same way. Hence, these approaches could be susceptible to bias if any participants receive antihypertensive treatment for some other reason, such as due to diabetes, CHD or migraine.

The approaches in the final group utilise all the observed data within an analysis, and apply a simple mathematical correction either to the modified BPs themselves (e.g. by adding a constant to any modified BP) or to the statistical likelihood function (e.g. by integrating over the normal probability density function to the right of any modified BP). These approaches assume that modified BPs are informative, and I therefore label this group the “Informative BP” group. The Informative BP approaches include: *Fixed Treatment Effect* (e), *Non-Parametric Adjustment* (i) and *Censored Normal Regression* (j). These approaches generally adjust for treatment in the same way for all individuals; thus, as we shall see in Section 1.3, they may be susceptible to bias if the effect of treatment varies between individuals.

1.1.5 Existing Work & Proposed Extensions

Previous work has compared and assessed the different approaches to correct analyses for the use of treatment either via simulation (Cook, 1997; Tobin *et al.*, 2005; McClelland *et al.*, 2008) or numerically (White *et al.*, 2003). These studies have investigated how the approaches perform under a number of conditions that aim to reflect different characteristics of real studies. For example, one such condition has been described as a *non-differential* intervention (White *et al.*, 2003). A non-differential intervention is where both the probability of receiving treatment and the effect of treatment depend only

upon the outcome of interest. Hence, for a study of BP, the intervention is non-differential if individuals only receive antihypertensives due to having high BP, and if the effects of any antihypertensive treatments depend only on BP.

Simulations have shown that, in general scenarios in which there is a non-differential intervention, all the approaches to analysis maintain approximately the correct levels of type I error (Tobin *et al.*, 2005; McClelland *et al.*, 2008). The approaches differ, however, in terms of how accurately they estimate the parameter coefficients, and in terms of the statistical powers obtained. For instance, *No Adjustment* (a), *Exclude* (b), and *Treatment as a Binary Covariate* (c) tend to underestimate the parameter coefficients and, thus, suffer reduced powers. *Binary Trait* (d) and *Median Method* (h) also appear to be low powered approaches. *Fixed Substitution* (f) and *Random Substitution* (g) sometimes perform well – with a high power and small bias in the parameter estimates – but on other occasions they perform poorly, i.e. yielding a lower power and a larger degree of bias. *Fixed Treatment Effect* (e), *Non-Parametric Adjustment* (i), and *Censored Normal Regression* (j) appear to be the best approaches; they typically yield a high power, and lead only to small degrees of bias in the estimates of the parameter coefficients. Possible explanations as to why each approach performs as such have already been discussed in Section 1.1.2.

A number of conditions have been tested in scenarios where the intervention is non-differential. For example, an “unobserved” covariate has been tested by simulating a factor that has a strong effect on BP but which is left unaccounted for in the analyses. A scenario in which some individuals without high blood pressure also receive treatment has also been tested. Furthermore, scenarios

have been tested in which the treatment effect is implemented in different ways. For example, the treatment effect has been simulated from a fixed normal distribution; as a percentage reduction of underlying BP; and by targeting a specific threshold for the observed phenotype to lower to. Generally, as reported in the literature (Tobin *et al.*, 2005; McClelland *et al.*, 2008), none of these conditions seems to result in any more than a small decrease in the statistical power for each approach, and a marginal increase in the bias of parameter estimates. However, some of the approaches have not been tested under each condition, and some of the conditions have not been individually tested in separate scenarios. As such, there remains a potential for further work in this area. Further work should clarify how the different approaches perform under each condition, and provide results that are comparable between scenarios. This is the focus of Section 1.2.

In contrast to the condition described as a “non-differential” intervention, an intervention is “differential” if the probability of receiving treatment and/or the effect of treatment depends on another factor (other than the outcome itself). For example, a differential threshold for receiving treatment would occur if antihypertensive medications are systematically prescribed to some participants (such as those with diabetes) at a lower threshold of BP than others (Raskin, 2003). Furthermore, a differential *treatment effect* would occur, for example, if the efficacy of treatment varies depending on genotype for a particular genetic variant (this is a type of *pharmacogenetic* interaction) (Turner *et al.*, 2001; Arnett *et al.*, 2009).

Currently, only a limited amount of work has investigated how the different approaches perform when there is a differential intervention. For instance, although White et al. (White *et al.*, 2003) examined the effects of both a differential threshold for receiving treatment and a differential treatment effect, they tested only *No Adjustment* (a), *Exclude* (b), and *Median Method* (h). Similarly, in addition to testing (a) and (b), a more recent study also tested *Treatment as a Binary Covariate* (c) and *Censored Normal Regression* (j); however, they only examined the effects of a differential probability for receiving treatment (McClelland *et al.*, 2008). Both of these studies found that the approaches to analysis behave very differently to one another under a differential intervention. Some of the approaches lead to bias under both conditions [e.g. (a)]; some of the approaches are only affected under one of the conditions [e.g. (b)]; and some seem to be unaffected in both conditions [e.g. (h)]. Where bias is obtained, this can impact upon both the power and type I error rate of an approach.

Given these findings, at present there remains no conclusive advice as to which approach to analysis to use in practice. Further work is therefore needed to clarify the most appropriate methods to use in different situations, such as the plausible scenario in which there is a differential intervention. This is the primary aim of this chapter. Section 1.2 focuses on investigating how the different approaches perform in different situations under a *non-differential* intervention. Building upon this, Section 1.3 then investigates how the approaches perform under a *differential* intervention. Each of the two subsequent sections performs a set of simulation studies focussing on cross-sectional genetic-association studies of BP. In particular, the aim of each

scenario is to assess the approaches described in Section 1.1.2 in terms of their ability to correct for the use of treatment, when the primary interest of each analysis is to estimate the marginal effect of a particular genetic variant (i.e. the effect of the variant unmitigated by treatment) on BP.

1.2. Simulation Studies: Non-differential intervention

This section assesses the approaches to correct for the use of treatment. Each approach is tested via simulation, and its capability to detect and estimate the marginal effects of two genetic variants on systolic blood pressure (SBP) is observed. As described below, one of these variants has a null effect and is used to estimate the type I error rate for each approach, while the other variant increases SBP, and is used to estimate power. The approaches are tested under a *non-differential intervention* in this section, where, as described by White et al. (White *et al.*, 2003), both the allocation of treatment and the effect of treatment depend solely on blood pressure.

The approaches are first tested in the *General Simulation Study*, which is a “baseline” scenario in which no particular model assumptions are contravened (see Section 1.2.1). This scenario, thus, demonstrates the potential levels of performance attainable by each approach. The approaches are then tested in several further scenarios that reflect different characteristics of real studies (Section 1.2.2-1.2.5). The General Simulation Study forms the basis of all subsequent scenarios, which are simulated by altering one or more of its properties. Hence, Section 1.2.1 provides a full description of the simulation

method, while subsequent sections detail only those properties that differ from the General Simulation Study.

Each scenario aims to estimate the marginal effect of a particular genetic variant on the underlying BP in the whole study population (i.e. the main effect of a SNP on the blood pressure that would have prevailed in the absence of antihypertensive treatment taken by a proportion of the population). Monte Carlo estimates of the statistical power, the type I error rate, and the mean level of bias are derived with respect to a single nucleotide polymorphism (SNP) for each approach described in Section 1.1.2. A SNP is a genetic variant that has one of two possible alleles on each of the two homologous chromosomes. For any particular SNP, the allele with the lowest frequency within a given population is known as the *minor allele*. Thus, for a particular SNP, an individual may have 0, 1 or 2 copies of the minor allele.

1.2.1 General Simulation Study

The General Simulation Study is designed to represent a population-based study of BP, consisting of 2,000 unrelated participants aged between 25 and 80 years. The following notation is consistent with that introduced in Section 1.1.2.

1.2.1.1 Simulation Method

For the i^{th} participant ($i = 1, \dots, 2000$), Z_i denotes underlying SBP (in mmHg); AGE_i denotes age (in years); SEX_i denotes sex (1 = male; 0 = female); g_{1i} and g_{2i} denote the genotypes for two independent SNPs, which have allele frequencies of 0.3 and which are centred for comparability across different SNP effect sizes (= 0, 1 or 2 copies of the minor allele for each SNP); and ε_i denotes

a normally distributed random error. AGE_i is generated from a uniform distribution with parameters 25 and 80; SEX_i is generated from a Bernoulli distribution with probability 0.5; and g_{1i} and g_{2i} are generated independently – each from two Bernoulli trials with probability 0.3.

For each individual, the underlying SBP, Z_i , is simulated from a linear regression model with an additive genetic effect (i.e. where 2 copies of the minor allele yields twice the effect of 1 copy):

Equation 12

$$Z_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 g_{1i} + \beta_4 g_{2i} + \varepsilon_i,$$

where $\beta_0 = 110$, $\beta_1 = 0.4$, $\beta_2 = 3$, $\beta_3 = 2$, $\beta_4 = 0$, and $\varepsilon_i \sim N(0,18)$. Note, thus, that the SNPs g_1 and g_2 are generated to evaluate power and type I error respectively. These simulation characteristics are chosen to generate a realistic distribution of SBP for a population-based genetic epidemiological study of blood pressure. The simulated effect size of g_1 is larger than would typically be expected, however, but is necessary here to ensure that the analysis models are adequately powered.

A condition for allocating treatment is implemented next. Following the definition outlined in the JNC VII (Chobanian *et al.*, 2003), an individual with underlying SBP greater or equal to 140 mm Hg is labelled as “hypertensive”, and will possibly receive treatment. In practice, not all hypertensive individuals receive antihypertensive medication; antihypertensives are therefore administered with probability 0.75. The binary variable $TREAT_i$ denotes treated status (1 = Yes; 0 = No), and is thus generated from a Bernoulli distribution with

probability 0.75. In the General Simulation Study, only *hypertensives* are administered treatment and, hence, $TREAT_i$ is always 0 if $Z_i < 140$.

For individual i , an *observed SBP*, Y_i , is generated to represent the BP measurements typically obtained within studies. For individuals who use antihypertensives, Y_i is derived by subtracting a treatment effect from Z_i . The size of the treatment effect is denoted γ_i , and γ_i is generated randomly from a normal distribution with mean 15 mmHg and variance 4^2 [i.e. $\gamma_i \sim N(15, 4^2)$]. To avoid any unrealistic cases where the treatment directly *increases* SBP, γ_i is truncated at 0.

Table 2 below summarises the simulation properties for the General Simulation Study.

Parameter	General Simulation Study
Sample Size: n	2000
Intercept: β_0	110
Age: AGE_i ; β_1	Uniform [25-80] ; 0.4
Sex: SEX_i ; β_2	Bernoulli (0.5) ; 3
Gene: g_{1i} ; β_3	Bin(2, 0.3) ; 2
Gene2: g_{2i} ; β_4	Bin(2, 0.3) ; 0
Random Error: ε_i	N(0, 18)
Hypertensive Criterion:	SBP \geq 140
$P\{TREAT_i=1 \text{hypertensive}\}$:	0.75
Treatment Effect : γ_i	N(15, 4^2)

Table 2: Simulation Properties for the General Simulation Study.

The General Simulation Study comprises 1,000 simulated datasets on each of which every approach to analysis described in Section 1.1.2 is performed. In

addition, a further approach is also performed – for comparison purposes – which analyses the underlying SBP for all subjects within each study. Although, in practice, the underlying SBP is not observable for all participants due to the use of antihypertensive treatments, this analysis of the underlying SBP highlights the optimal level of performance reasonably attainable given the simulation characteristics (i.e. the simulated SNP effect size, the sample size, etc).

For each approach, Monte-Carlo estimates of the effects of the model parameters *AGE*, *SEX*, g_1 and g_2 are derived, as well as estimates of coverage (i.e. number of times the 80% confidence interval around a parameter estimate contains the true value), and the power and type I error rates (with respect to the genetic variants g_1 and g_2 accordingly).

Some of the approaches to analysis require particular values to be specified [such as the constant c for approach (e), and the constant m for approach (f)]. Guidance for selecting these imputation values has typically been provided by the original proposers (e.g. in (White *et al.*, 1994; Cui *et al.*, 2002; Hunt *et al.*, 2002; Cui *et al.*, 2003; White *et al.*, 2003)), and depends on knowledge about the specific antihypertensive drugs under consideration. A number of different values can potentially be used with each approach, however, and the simulations that follow use a range of illustrative parameter values. Table 3 overleaf provides a brief summary of each approach and lists the parameter values used in these simulation studies.

Methods	Description	Parameter Values
Naïve:		
(a) No Adjustment	Ignore use of treatment; analyse all observations in a linear regression model.	
(b) Exclude	Exclude any participants who use antihypertensive medication from the analysis.	
(c) Treatment as a Binary Covariate	Adjust for antihypertensive treatment use by fitting $TREAT_i$ as a binary covariate.	$TREAT_i = 1$ if individual i uses antihypertensive medication; $TREAT_i = 0$ otherwise.
Substitution:		
(d) Binary Trait	Define a binary “hypertension” outcome, and fit a logistic regression model to the data.	$hypertension_i = 1$ if $TREAT_i = 1$ or if $Z_i \geq 140$ mmHg; $hypertension_i = 0$ otherwise.
(f) Fixed Substitution	Substitute <i>modified BPs</i> with the constant m .	$m = 130$ mmHg; and 140 mmHg.
(g) Random Substitution	Substitute <i>modified BPs</i> with values generated randomly from a normal distribution.	$\sim N(150, 5^2)$ truncated to [140,160]
(h) Median Method	Substitute <i>modified BPs</i> with the value k , and fit a quantile regression to the data.	$k = 140, 150, 160$ in the General Simulation Study; $k = 160, 180$ and 200 in subsequent scenarios.
Informative BP:		
(e) Fixed Treatment Effect	Add the constant c to <i>modified BPs</i> .	$c = 5; 10; \text{ and } 15$.
(i) Non-parametric Adjustment	Apply an algorithm to derive a set of adjusted residuals; adjust <i>modified BPs</i> by adding the difference between the current adjusted and raw residuals.	
(j) Censored Regression	Assume that <i>modified BPs</i> are right-censored; fit a Tobit-model to the data.	

Table 3: Summary of the analysis models with parameter values used in the simulated studies. Shaded and non-shaded regions denote the three categories of approaches: Naïve, Substitution, and Informative BP.

Descriptive statistics for the General Simulation Study are presented in Table 4 below. Note that the mean proportion of individuals who receive antihypertensives within these studies is relatively high (27.87%). This was deliberately simulated at a relatively high level in order to better illustrate the potential implications of failing to adequately handle *modified BPs*. Note, however, that real examples of studies with a similar proportion of individuals on treatment do exist (such as in studies of older populations, e.g. (Wang *et al.*, 2007a)).

Summary Statistics	General Simulation Study
Mean Underlying SBP (S.D.)	133.71 (19.2)
for <i>treated</i> subjects	153.24 (10.4)
for <i>untreated</i> subjects	126.17 (16.2)
Mean Observed SBP (S.D.)	129.53 (16.0)
for <i>treated</i> subjects	138.24 (11.2)
for <i>untreated</i> subjects	126.17 (16.2)
% SBP>140	38.17
% SBP>150	19.83
% SBP>160	8.56
% Treated	27.87
Mean Treatment Effect (SD)	-15.00 (4.0)

Table 4: Descriptive Statistics for 1,000 datasets of the General Simulation Study.

1.2.1.2 Results

As noted earlier, the analyses focus on estimating the marginal effect of the genetic variants g_1 and g_2 on BP. All approaches fit the model in Equation 1 [except for *Treatment as a Binary Covariate* (c) and *Binary Trait* (d)] by replacing the underlying SBP, Z_i , with the values imputed according to each method (e.g. for *No Adjustment* (a), these would simply be the observed SBPs,

Y_i for all individuals). For method (c), Equation 1 is fitted with the observed SBPs, Y_i as outcome and an additional binary covariate for treatment, $TREAT_i$, is included on the right hand side. For method (d), a logistic regression model is fitted to the dichotomous outcome, $hypertension_i$ (see Table 3).

The full results for the General Simulation Study are presented in Table 33 in Appendix A, and a summary of the results is presented graphically in Figure 1 below. In Figure 1, the statistical power to detect the marginal effect of g_1 and the type I error rate with respect to the null effect of g_2 are shown on the left-vertical axis (at the 5% level of significance). The mean bias with respect to the estimated coefficient of g_1 (only), with standard error, is also shown (in mmHg) – on the right-vertical axis. Note that any references to mean bias in the following text refer to this estimated coefficient of g_1 . Note, furthermore, that the measures of power, type I error and bias depend on factors such as the sample size, the minor allele frequency of the SNP of interest, the size of the SNP effect, the proportion of individuals treated within each study, and the magnitude of the treatment effect, etc.

The approaches to analysis are arranged in classes across the x-axis, but, for reasons that will be demonstrated in Section 1.3, not necessarily in the order introduced in Section 1.1.2. Results for the additional analysis of underlying SBP are also included on the far left of the plot.

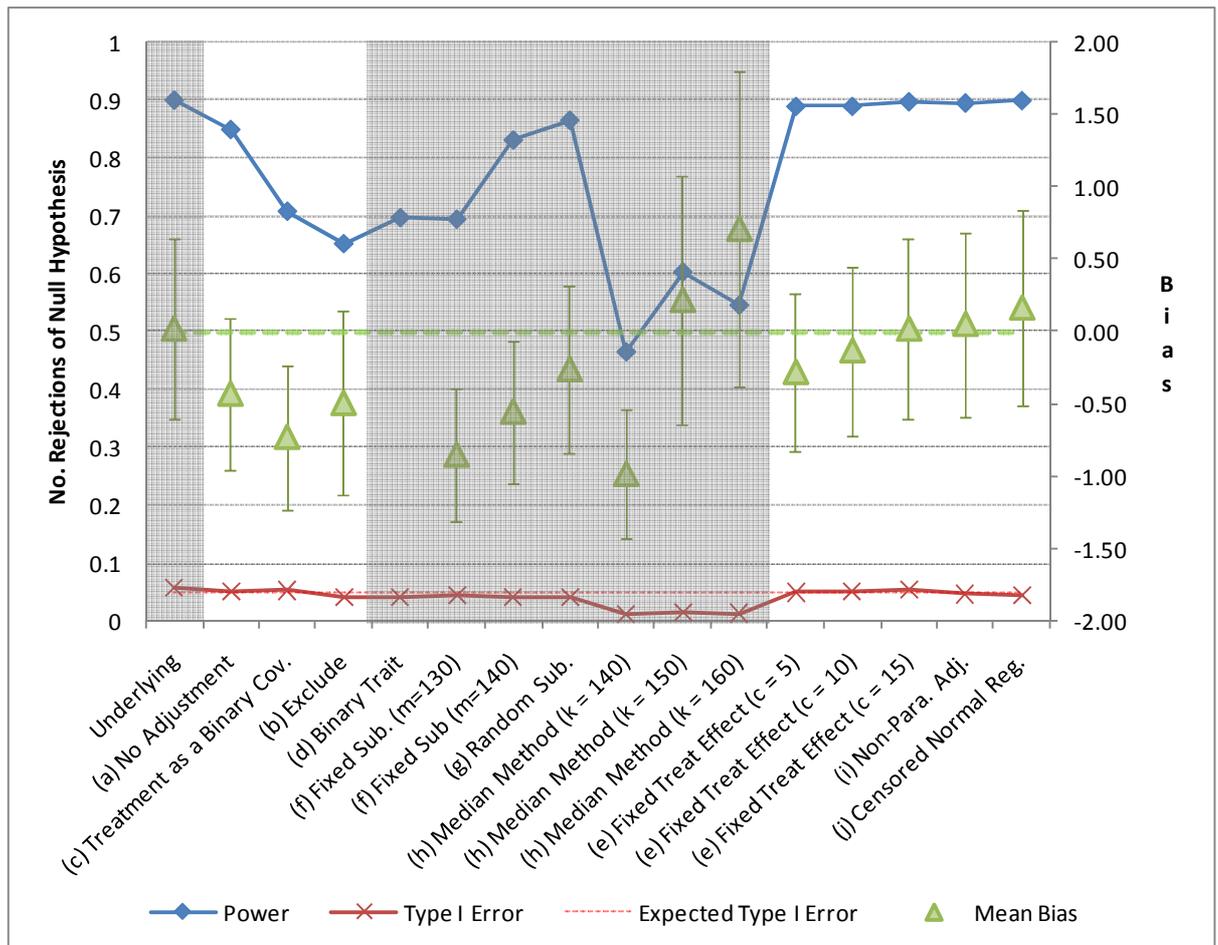


Figure 1: Graphical representation of results for the General Simulation Study. Approaches are arranged here in categories (from left to right): naïve, substitution, informative phenotype. Power [relative to the g_1 coefficient, β_3] and Type I Error (relative to the g_2 coefficient, β_4) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

Focussing firstly on the analysis of underlying SBP, the power obtained for detecting the genetic variant g_1 is 0.9. This is the only approach unaffected by any bias due to the use of treatment and, hence, a power of 0.9 seems the maximum reasonably attainable by any approach. Other aspects of the results for this analysis are consistent with those expected of a conventional and complete analysis (i.e. it yields approximately correct coverage and type I error rates, with only a small degree of bias). As Table 33 in Appendix A shows, the coverage for this approach is also correct.

Results for the other approaches are most easily considered in terms of the three classes described in Section 1.1.4. For the three Naïve approaches [*No Adjustment* (a), *Exclude* (b) and *Treatment as a Binary Covariate* (c)], estimates of the effects of all the covariates are shrunken towards the null (i.e. closer to the null effect of zero), and there are consequent losses of statistical power. For example, where the simulated β_3 value is 2, the mean estimated values, $\hat{\beta}_3$, range between 1.27 and 1.57 (see Table 33, Appendix A), and the powers range between 0.65 and 0.85. Hence, although the Naïve approaches retain the correct type I error rates, they are clearly suboptimal methods of analysis.

With regards to the Substitution approaches [*Binary Trait* (d), *Fixed Substitution* (f), *Random Substitution* (g), and *Median Method* (h)], the results seem similarly unfavourable. For instance, although the parameter coefficients for *Binary Trait* (d) cannot be compared to the other approaches because they are log-odds ratios (obtained by logistic regression), (d) has a very low statistical power ($\approx 70\%$). Likewise, although *Random Substitution* (g) yields a reasonable power ($\approx 87\%$), a comparison with the similar *Fixed Substitution* (f) approach (power $\approx 70\% - 83\%$) shows that (f) and (g) are highly sensitive to the “substitution values” used in place of any *modified BPs*. In these simulations, the value of 130 mmHg [i.e. which is used as a substitution value for (f)] is known to be below the threshold for initiating antihypertensive treatment; however, in practice, the relevant threshold is not always known and even when guidelines for treatment are widely available they are not always rigidly adhered to (Wolf-Maier *et al.*, 2003). As the substitution value is increased from 130 mmHg to 140 mmHg [with (f)] and to the mean value of 150 mmHg [with (g)], the mean bias decreases from -0.9 mmHg to -0.3 mmHg. Hence, although (f) and (g) can

potentially perform reasonably, they are highly influenced by the choice of substitution value.

The *Median Method* (h) also appears sensitive to its substitution parameter (i.e. the constant k). When k is 140, the g_1 effect is underestimated (mean bias ≈ -1 mm Hg), and when k is 150 or 160, the g_1 effect is overestimated (mean bias ≈ 0.2 to 0.7 mmHg). White *et al.* (White *et al.*, 2003) state that (g) should actually be *insensitive* to k so long as the value chosen for k is sufficiently large. Subsequent sections therefore use greater values for k instead of the three values used here. Nevertheless, in this scenario, with any of the three seemingly plausible values of k tested, (g) yields a low statistical power ($\approx 47\%$ - 60%) and a lower level of type I error than expected ($\approx 1.5\%$), despite the fact that there were never more than 50% of individuals on treatment in any simulation run.

The results that most closely resemble those of the analysis of underlying SBP are obtained by the Informative BP approaches [*Fixed Treatment Effect* (e), *Non-parametric adjustment* (i) and *Censored Normal Regression* (j)]. Each of the Informative BP approaches yields a high statistical power close to 90%, the expected type I error rate, and only a small magnitude of bias [mean bias ≈ -0.25 to 0.2 mm Hg]. These analyses therefore seem reasonable approaches to correct for the use of antihypertensives. For *Fixed Treatment Effect* (e), the results obtained are relatively stable with the different values of c tested (i.e. where c is the constant added to *modified BPs* to reverse the negative effect of treatment) – with decreasing levels of bias as c is closer to the simulated treatment effect. In agreement with previous findings (Tobin *et al.*, 2005),

approach (e) thus seems relatively insensitive to the different choices of c used here, and performs well even when c differs considerably from the simulated treatment effect.

1.2.2 Scenario 1: Unobserved Covariate

Scenario 1 aims to test the approaches under a condition where the simulation model is different from the model used in the analyses. In reality, the true model is unknown, and factors may exist that influence the phenotype but are not measured during a study. An additional covariate is therefore simulated in this scenario that is left unaccounted for in the analyses. This is referred to as an “unobserved covariate”. An unobserved covariate could represent factors such as dietary salt intake or the use of oral contraceptives, which may affect SBP but are not always recorded in real studies. Although the simulated “*random error*” term might also account for these factors, the unobserved covariate implemented here asserts a more systematic influence upon BP.

1.2.2.1 Simulation Method

The unobserved covariate, u_i , is implemented in the present scenario with a prevalence of 0.2 (1 = exposed; 0 = unexposed), and the simulation model is therefore

$$Z_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 g_{1i} + \beta_4 g_{2i} + \beta_5 u_i + \varepsilon_i,$$

where $\beta_0 = 108$, $\beta_5 = 10$, and all other properties are as before (see Section 1.2.1.1). The full simulation properties are summarised in Table 5 below. Because the constant value is adjusted in this scenario compared to the

General Simulation Study (i.e. 108 Vs 110 mmHg respectively), the unobserved covariate implemented here does not affect the simulated distribution of SBP. Hence, the descriptive statistics in this scenario are approximately the same as those from the General Simulation Study (shown in Table 4), and are not provided.

Parameter	Scenario One
Sample Size: n	2000
Intercept: β_0	108
Age: age_i ; β_1	Uniform [25-80] ; 0.4
Sex: sex_i ; β_2	Bernoulli (0.5) ; 3
Gene: g_{1i} ; β_3	Bin(2, 0.3) ; 2
Gene2: g_{2i} ; β_4	Bin(2, 0.3) ; 0
Unobserved Covariate: u_i ; β_5	Bernoulli (0.2); 10
Random Error: ε_i	N(0, 18)
Hypertensive Criterion:	SBP \geq 140
$P\{ TREAT_i=1 \text{hypertensive}\}$:	0.75
Treatment Effect : γ_i	N(15, 4 ²)

Table 5: Simulation Properties for Scenario 1.

1.2.2.2 Results

Figure 2 below presents a graphical summary of the results for Scenario 1 and Table 34 in Appendix A shows the full table of results. As can be seen, only small changes in the performance of the approaches are obtained here compared with in the General Simulation Study. For instance, the analysis of underlying SBP yields a very small decrease in statistical power (power = 0.886), and the other approaches therefore also reflect this.

Overall, the general pattern of results in this scenario is the same as in the General Simulation Study. The Naïve approaches [*No Adjustment* (a), *Exclude* (b), and *Treatment as a Binary Covariate* (c)] perform poorly, with low powers (power $\approx 0.6 - 0.85$) and parameter estimates shrunken to the null (mean $\hat{\beta}_3 \approx 1.25 - 1.6$). *Binary Trait* (d) also yields a low statistical power ($= 0.662$), while *Fixed Substitution* (f) and *Random Substitution* (g) are sensitive to the substitution parameters [e.g. the constant m for (f)]. For the *Median Method* (h), greater values of k are now used (i.e. $k = 160, 180$ and 200 in this scenario), and the results approximately converge to the same mean parameter estimates irrespective of the three choices of k – as expected. Nevertheless, (h) again yields a low statistical power with all values of k tested (power ≈ 0.50), as well as drastically overestimated parameter coefficients (e.g. mean $\hat{\beta}_3 \approx 2.61$).

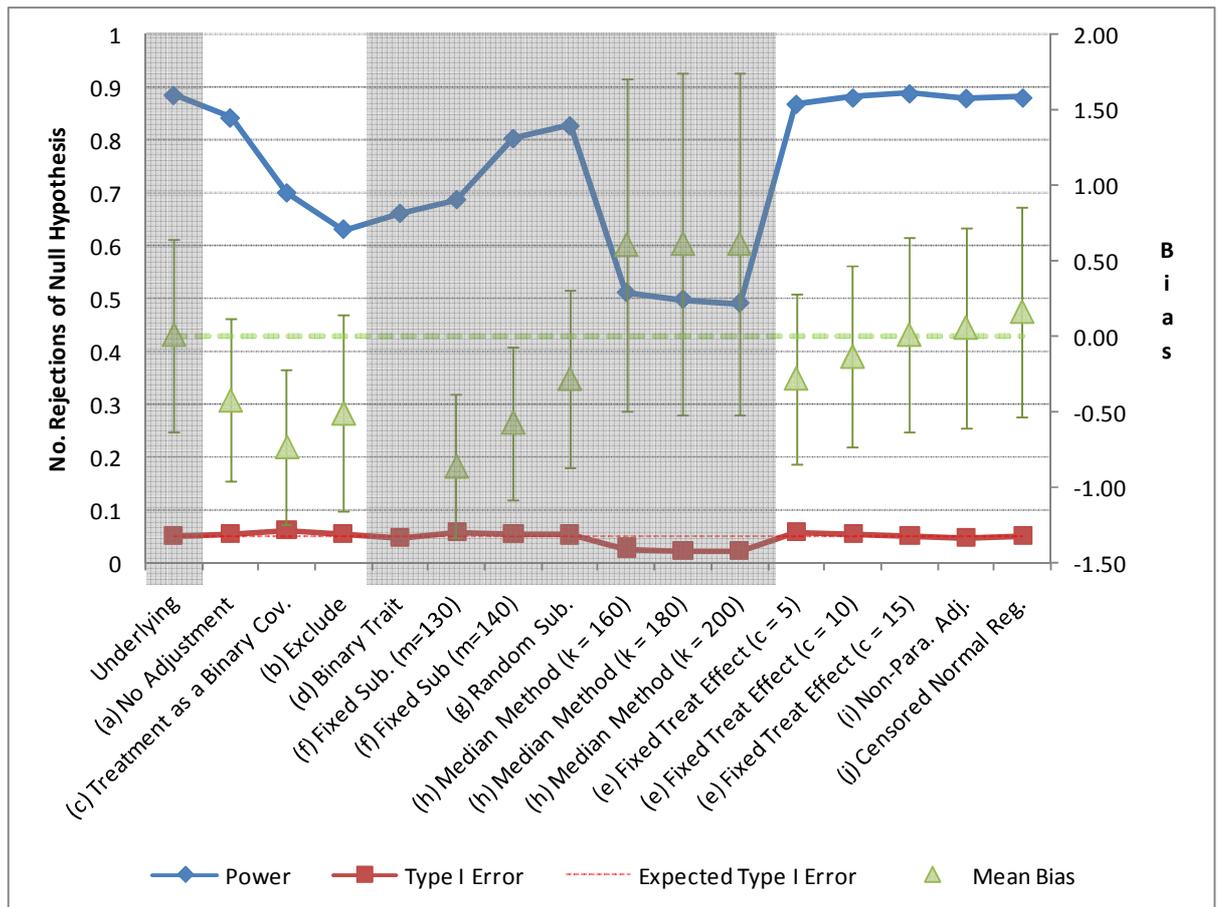


Figure 2: Graphical representation of the results for Scenario 1. Approaches are arranged in the categories (from left to right): Naïve, Substitution, Informative BP. Power (relative to the β_1 coefficient, β_3) and Type I Error (relative to the β_2 coefficient, β_3) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

The Informative BP approaches [*Fixed Treatment Effect (e)*, *Non-Parametric Adjustment (i)*, and *Censored Normal Regression (j)*] again perform best, with results closely resembling those obtained by the analysis on underlying SBP. The Informative BP approaches yield high powers (≈ 0.87 - 0.89), and again yield only small magnitudes of bias (e.g. mean $\hat{\beta}_3 \hat{\sigma}_1 \approx 1.73 - 2.16$). It would thus seem that the most appropriate methods for analysis, in the case where an unobserved covariate is suspected, are the Informative BP approaches.

1.2.3 Scenario 2: Treated Normotensives

Scenario 2 aims to simulate the usage of antihypertensives for other reasons than lowering blood pressure. For instance, beta-blockers may be used to treat conditions such as migraine and glaucoma in addition to hypertension. As such, the use of antihypertensives is not restricted just to individuals with high blood pressure. Individuals with lower blood pressure can also be prescribed antihypertensive treatments, and this scenario investigates how the approaches perform when the simulated studies include a number of these individuals.

1.2.3.1 Simulation Method

This scenario allocates treatment to 60% of individuals with SBP of at least 140mmHg (*hypertensives*), and additionally to 10% of individuals with SBP below 140mmHg (*normotensives*). This is in contrast to the General Simulation Study, which allocates treatment to hypertensives only, with a probability of 75%. Nevertheless, as the allocation of treatment again solely depends on BP here, the intervention remains *non-differential* in this scenario.

For consistency with other scenarios, all simulation properties other than the above condition (i.e. for allocating treatment), are the same in this scenario compared to the General Simulation Study. The simulation properties for Scenario 2 are listed in Table 6 below.

Parameter	Scenario Two
Sample Size: n	2000
Intercept: β_0	110
Age: age_i ; β_1	Uniform [25-80] ; 0.4
Sex: sex_i ; β_2	Bernoulli (0.5) ; 3
Gene: g_{1i} ; β_3	Bin(2, 0.3) ; 2
Gene2: g_{2i} ; β_4	Bin(2, 0.3) ; 0
Random Error: ϵ_i	N(0, 18)
Hypertensive Criterion:	SBP \geq 140
$P\{TREAT_i=1 hypertensive\}$:	0.6
$P\{TREAT_i=1 normotensive\}$:	0.1
Treatment Effect : γ_i	N(15, 4 ²)

Table 6: Simulation properties for Scenario 2.

Table 7 below shows descriptive statistics for Scenario 2. As can be seen, this scenario yields similar descriptive statistics to the General Simulation Study, with approximately the same overall percentage of individuals on treatment (\approx 28%). However, around 22% of the treated individuals are *normotensive* here, and, hence, the overall mean SBP (both underlying and observed) for treated individuals is lower (i.e. mean observed SBP for treated subjects = 131.38 mmHg here Vs 138.24 mmHg in the General Simulation Study).

Summary Statistics	Scenario Two
Mean Underlying SBP (S.D.)	133.70 (19.2)
for <i>treated</i> subjects	146.39 (16.9)
for <i>untreated</i> subjects	128.63 (17.6)
Mean Observed SBP (S.D.)	129.41 (17.6)
for <i>treated</i> subjects	131.38 (17.4)
for <i>untreated</i> subjects	128.63 (17.6)
% SBP>140	37.17
% SBP>150	19.80
% SBP>160	8.55
% Treated	28.00
% Treated Individuals who are <i>Normotensive</i>	21.98
Mean Treatment Effect (SD)	-15.00 (4.0)

Table 7: Descriptive statistics for Scenario 2 (based on 1,000 simulation runs).

1.2.3.2 Results

Figure 3 below summarises the results for Scenario 2 graphically, and Table 35 in Appendix A shows the full table of results. Before reflecting upon the results for this scenario, it is worthwhile to consider how the different classes of approach might be expected to perform here. For instance, the Substitution approaches [*Binary Phenotype* (d), *Fixed Substitution* (f), *Random Substitution* (g), and *Median Method* (h)] assume that any individual who uses antihypertensive treatment is hypertensive, and they replace the BP measurements corresponding to these individuals with alternative values typical of hypertension. As such, these approaches misclassify any treated normotensives in this scenario, and may be expected to be biased. In contrast, neither the Naïve nor the Informative BP approaches make any specific assumptions regarding the use of treatment. Hence, there is no reason to

suspect that these approaches will be adversely affected by the condition implemented in this scenario.

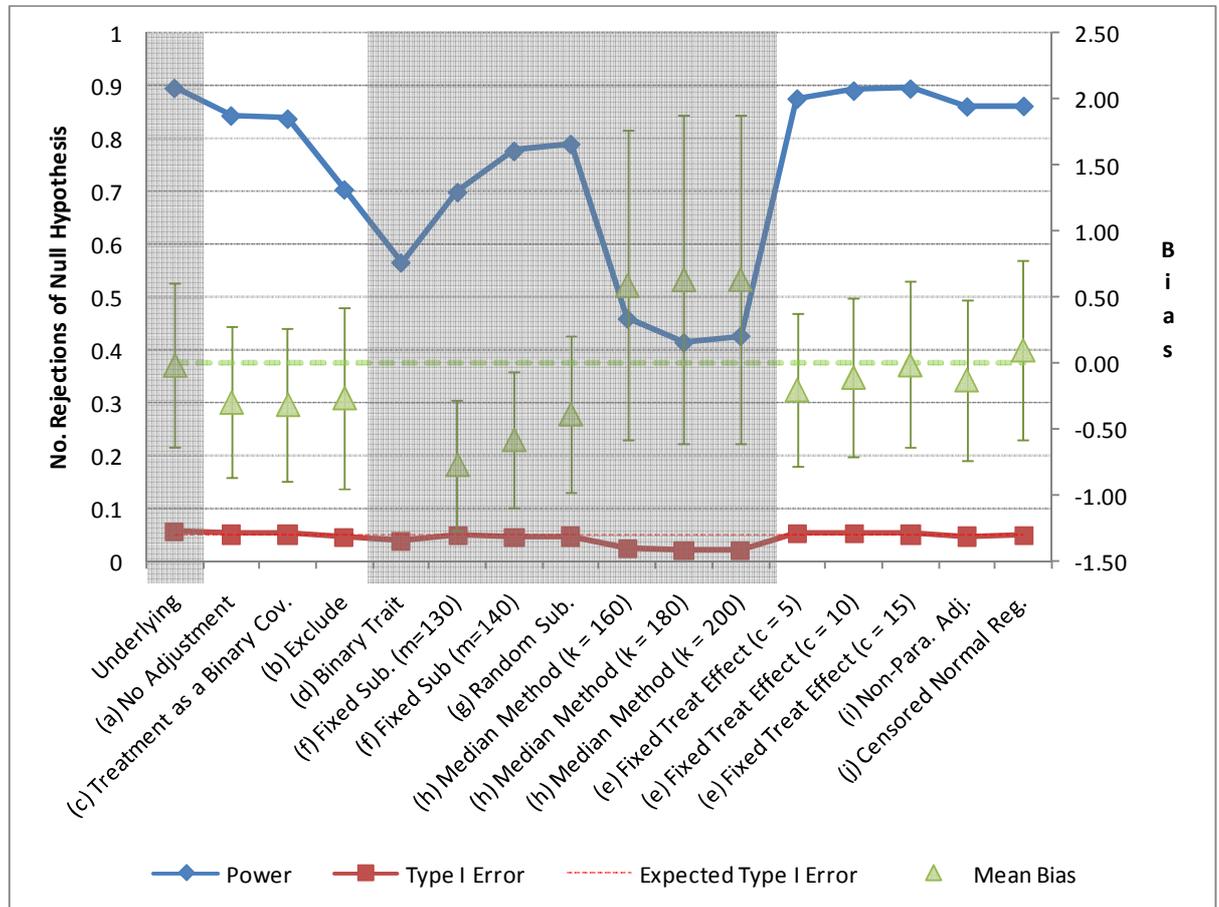


Figure 3: Graphical representation of results for Scenario 2. Approaches are arranged in the categories (from left to right): Naive, Substitution, Informative BP. Power (relative to the g_1 coefficient, β_3) and Type I Error (relative to the g_2 coefficient, β_3) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

As Figure 3 shows, the results obtained in this scenario reflect the predictions made. For instance, the Substitution approaches all perform worse here than in the General Simulation Study. The powers for these approaches are lower ($\approx 0.4 - 0.8$), and the bias in the estimates of the parameter coefficients is increased [e.g. mean $\hat{\beta}_3 \approx 1.24 - 1.62$ for (f) and (g)]. For the Informative BP approaches, the results closely resemble those in the General Simulation

Study. Although (i) and (j) yield slightly lower powers (≈ 0.86), they yield no major increase in the bias of the parameter estimates (e.g. mean $\hat{\beta}_3 \approx 1.88$ and 2.11 respectively).

For the Naïve approaches, *Exclude* (b) and *Treatment as a Binary Covariate* (c) actually perform better here than in the General Simulation Study, while *No Adjustment* (a) performs similarly. Approaches (b) and (c) yield greater powers here compared with the General Simulation Study [power = 0.70 and 0.84 respectively], and all three Naïve approaches yield smaller magnitudes of bias (e.g. mean $\hat{\beta}_3 \approx 1.69 - 1.74$). These improvements can be explained in one of two ways. Firstly, as this scenario administers treatment to some normotensives in addition to a proportion of the hypertensives, any bias due to treatment becomes more balanced between those with high and those with low BP. Secondly, the improvement for (c) may be explained by noting the magnitude of the difference in observed SBP between treated and untreated subjects here (i.e. mean difference ≈ 2.8 mmHg). As this difference is small in this scenario, the act of adjusting for differences between treated and untreated subjects (i.e. by modelling treatment as a covariate) actually achieves very little. *Treatment as a Binary Covariate* (c) therefore behaves similarly to *No Adjustment* (a) in this scenario.

Despite the Naïve approaches' improved performance here, it should be noted that the Informative Phenotype approaches still perform better. Similarly, the improvements observed here for the Naïve approaches will not necessarily replicate in other situations, which, for example, may have different distributions of SBP.

1.2.4 Scenario 3: Combination Therapy

In practice, it is often necessary to prescribe more than one class of treatment (i.e. *combination therapy*) or a higher dosage of treatment to individuals with especially high blood pressure (BP). The magnitude of BP lowering due to treatment is therefore increased for these individuals, in order to lower BP to a safer level. This scenario simulates the use of combination therapy, with the aim of assessing how the approaches perform when the distribution of the treatment effect differs between individuals.

1.2.4.1 Simulation Method

Individuals initially allocated to receive treatment are administered a second treatment, if, after the first treatment, the observed SBP remains above 140 mmHg. The effect of the second treatment is implemented in the same way as the original treatment, i.e. it is generated randomly from a normal distribution with a mean of 15 and variance 4 (and truncated at zero).

Table 8 below summarises the simulation properties for this scenario and Table 9 presents the descriptive statistics.

Parameter	Scenario Three
Sample Size: n	2000
Intercept: β_0	110
Age: age_i ; β_1	Uniform [25-80] ; 0.4
Sex: sex_i ; β_2	Bernoulli (0.5) ; 3
Gene: g_{1i} ; β_3	Bin(2, 0.3) ; 2
Gene2: g_{2i} ; β_4	Bin(2, 0.3) ; 0
Random Error: ϵ_i	N(0, 18)
Hypertensive Criterion:	SBP ≥ 140
$P\{TREAT_i=1 \text{hypertensive}\}$:	0.75
1 st Treatment Effect : γ_i	N(15, 4 ²)
Allocation to 2 nd Treatment	If $TREAT_i=1$ & Observed SBP ≥ 140
2 nd Treatment Effect: γ_{2i}	N(15, 4 ²)

Table 8: Simulation properties for Scenario 3.

Table 9 shows that 37.5% of the participants originally allocated to receive treatment also receive a second treatment (approx. 10% of the subjects overall). Subjects treated twice have a mean underlying SBP approximately 16 mmHg greater than that for subjects treated just once (approx. 163 mmHg Vs 147 mmHg), but the difference in the mean *observed* SBP between these two subgroups is only 4 mmHg (approx. 135 mmHg Vs 131 mmHg). As before, the mean observed SBP for subjects not allocated to treatment (≈ 126 mmHg) is lower than the mean in both treatment groups, but the differences here between treated and untreated individuals are smaller, on average, than in the General Simulation Study.

Summary Statistics	Scenario Three
Mean Underlying SBP (S.D.)	133.68 (19.2)
for <i>untreated</i> subjects	126.15 (16.2)
for subjects <i>treated once</i>	147.04 (4.8)
for subjects <i>treated twice</i>	163.57 (9.0)
Mean Observed SBP (S.D.)	127.95 (14.6)
for <i>untreated</i> subjects	126.15 (16.2)
for subjects <i>treated once</i>	131.34 (5.1)
for subjects <i>treated twice</i>	134.87 (9.4)
% Treated	27.80
% Treated Twice	10.45
% Treated Twice Treated	37.50
Mean Treatment Effect for Subjects Treated Once (SD)	-15.80 (3.8)
Mean Treatment Effect for Subjects Treated Twice (SD)	-28.67 (5.6)
Mean Treatment Effect (SD)	-20.56 (7.7)
% SBP>140	37.09
% SBP>150	19.78
% SBP>160	8.53

Table 9: Descriptive statistics for Scenario 3.

1.2.4.2 Results

Results for Scenario 3 are presented graphically in Figure 4 below, and the full results are shown in Table 36 in Appendix A. Before consideration of the results, however, it should be noted that several of the approaches will not be affected here. The Substitution approaches [i.e. Binary *BP* (d), *Fixed Substitution* (f), *Random Substitution* (g), and *Median Method* (h)] and *Exclude* (b) either disregard or substitute any observations corresponding to subjects on treatment, and therefore will not be influenced by any changes regarding the treatment. These approaches, thus, should yield identical results in this scenario compared with in the General Simulation Study.

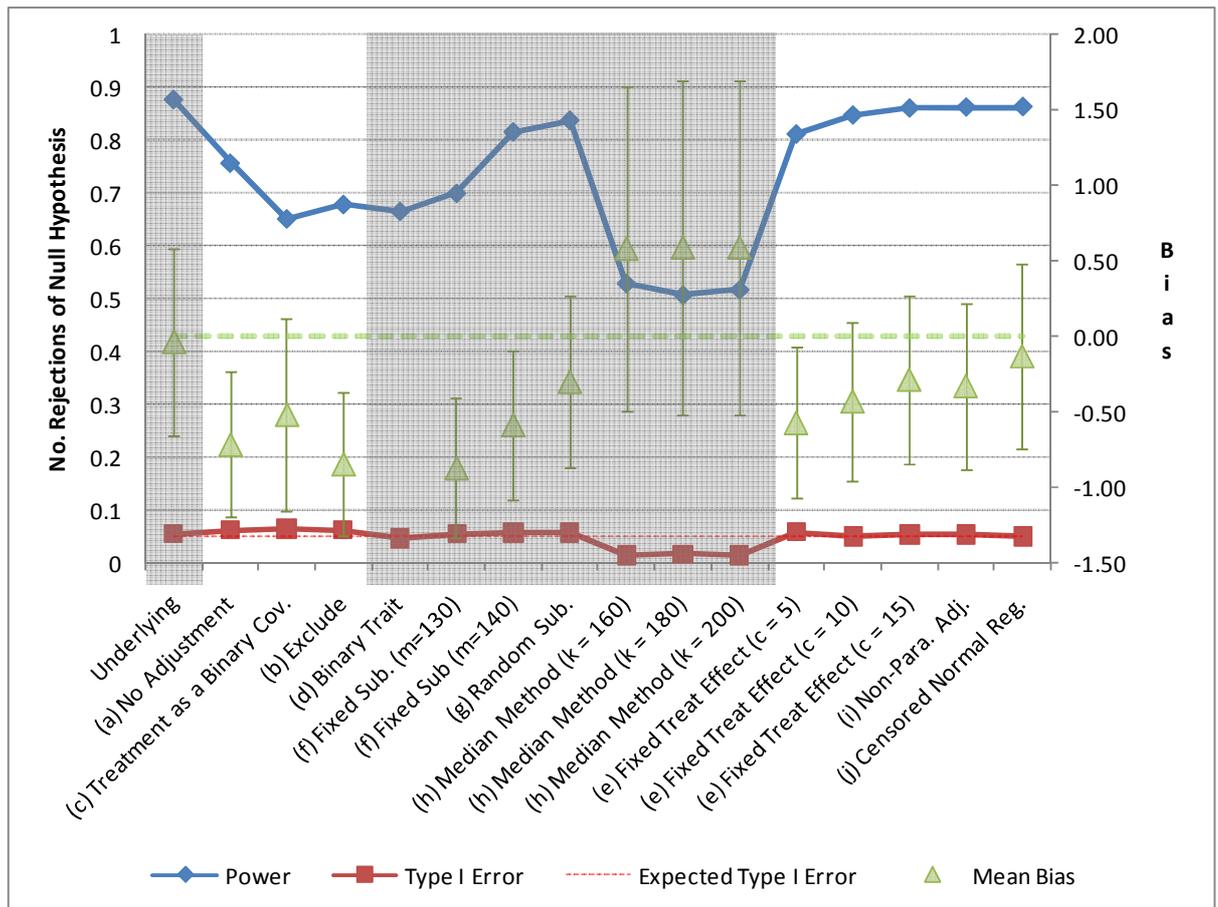


Figure 4: Graphical representation of results for Scenario 3. Approaches are arranged in the categories (from left to right): naïve, substitution, informative phenotype. Power (relative to the g_1 coefficient, β_3) and Type I Error (relative to the g_2 coefficient, β_4) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

As expected, (b), (d), (f), (g) and (h) all yield comparable results here compared to the General Simulation Study. However, as noted in Section 1.2.1, none of these approaches seem favourable compared to the Informative BP approaches. They generally have low powers to detect the genetic variant g_1 [e.g. power $\approx 0.50 - 0.70$ for (b), (d), and (h)] and yield estimates of the parameter coefficients shrunk to the null [e.g. mean $\hat{\beta}_3 \approx 1.1 - 1.7$ for (b), (f) and (g), mean $\hat{\beta}_3 \approx 2.6$ for (h)].

With the exception of *Treatment as a Covariate* (c), the other approaches perform slightly worse in this scenario compared with the General Simulation Study. The Informative BP approaches [*Fixed Treatment Effect* (e), *Non-Parametric Adjustment* (i), and *Censored Normal Regression* (j)] and *No Adjustment* (a) yield small decreases in power here, and marginally greater increases in bias [e.g. $\text{mean}\hat{g}_1 = 1.29$ for (a), and $\text{mean}\hat{g}_1 \approx 1.43 - 1.87$ for (e), (i) and (j)]. For (e), it is relatively straightforward to see why these problems occur. For instance, (e) imputes a fixed size treatment effect, which is inappropriate for those subjects on combination therapy. The mean treatment effect for individuals treated twice (-28.67mmHg) is greater than the three values of c tested, and the SBPs for these individuals are therefore underestimated. The shrinkage bias obtained in this scenario for (i) and (j) can be explained similarly. Approaches (i) and (j) make no distinction in the analyses between individuals treated once and individuals treated twice, and both subgroups have similar mean observed SBP (approx. 135 mmHg Vs 131 mmHg). The imputed SBPs, X_i , will therefore be similar for both subgroups despite large differences in the *underlying SBP*, on average, between individuals in the two subgroups (mean underlying SBP ≈ 163 mmHg Vs 147 mmHg). Hence, X_i underestimates Z_i for the individuals treated with “combination therapy”.

1.2.5 Scenario 4: Proportional Treatment Effect

Each of the scenarios so far has simulated the treatment effect by randomly sampling it from a fixed normal distribution. However, previous work has also simulated a treatment effect as a proportional reduction of the underlying phenotype (McClelland *et al.*, 2008). This results in individuals with higher BP

facing a greater reduction in BP due to treatment, and those with less extreme BP facing smaller treatment effects. This scenario assesses how the approaches perform when the treatment effect is applied as a proportional reduction of the underlying SBP.

1.2.5.1 Simulation Method

As in earlier scenarios, this scenario allocates treatment to hypertensives only, with a probability of 75%. Where applicable, treatment effects are applied by reducing an individual's underlying SBP by a proportion of the underlying SBP. The proportional reduction of SBP due to treatment varies between individuals, and, as in previous scenarios, is generated from a normal distribution. Four "sub-scenarios" are tested, each using a different distribution for generating the proportional reduction of SBP due to treatment. For instance, each sub-scenario generates the treatment effect, γ_i , from the following normal distributions (mean, SD^2): (0.05, 0.025^2), (0.1, 0.025^2), (0.15, 0.025^2), and (0.2, 0.023^2) respectively. Note that these distributions are chosen to maintain a similar variance for the treatment effect as in the General Simulation Study. Thus, the variance in the fourth sub-scenario differs from the others. Note, also, that each distribution is truncated at zero to prevent the treatment from ever directly increasing SBP.

Table 10 below lists the simulation properties for Scenario 4, which are identical to the General Simulation Study other than in terms of the implementation of the treatment effect.

Parameter	Scenario Four
Sample Size: n	2000
Intercept: β_0	110
Age: age_i ; β_1	Uniform [25-80] ; 0.4
Sex: sex_i ; β_2	Bernoulli (0.5) ; 3
Gene: g_{1i} ; β_3	Bin(2, 0.3) ; 2
Gene2: g_{2i} ; β_4	Bin(2, 0.3) ; 0
Random Error: ε_i	N(0, 18)
Hypertensive Criterion:	SBP \geq 140
$P\{TREAT=1 \text{hypertensive}\}$:	0.75
Treatment Effect (% Reduction of Underlying SBP)	N(5, 2.5)
	N(10, 2.5)
	N(15, 2.5)
	N(20, 2.3)

Table 10: Simulation Properties for Scenario 4.

Table 11 presents the descriptive statistics for this scenario. The only difference between the four sub-scenarios relates to the size of the treatment effect in each. For example, the treatment effect, on average, ranges between approx. -7.5 mmHg and -30 mmHg across the four sub-scenarios. The second sub-scenario – in which the mean proportional reduction in SBP due to treatment is 10% - is the most similar to previous scenarios, as the mean treatment effect here is -15.33 mmHg. The fourth sub-scenario is the most extreme, with a mean treatment effect of -30.65 mmHg. This sub-scenario is the only scenario simulated thus far in which the mean observed SBP for treated individuals is less than that for non-treated individuals.

Summary Statistics	Scenario Four			
	Mean % Reduction of SBP Due to Treatment			
	5	10	15	20
Mean Underlying SBP (S.D.)	133.70 (19.2)	133.69 (19.2)	133.72 (19.2)	133.72 (19.2)
for <i>treated</i> subjects	153.22 (10.4)	153.24 (10.5)	153.25 (10.5)	153.23 (10.4)
for <i>untreated</i> subjects	126.16 (16.2)	126.15 (16.2)	126.18 (16.2)	126.17 (16.2)
Mean Observed SBP (S.D.)	131.56 (17.2)	129.42 (15.7)	127.32 (14.8)	125.17 (14.7)
for <i>treated</i> subjects	145.53 (10.6)	137.91 (10.1)	130.27 (9.7)	122.58 (9.1)
for <i>untreated</i> subjects	126.16 (16.2)	126.15 (16.2)	126.18 (16.2)	126.17 (16.2)
% SBP>140	37.13	37.13	37.19	37.19
% SBP>150	19.78	19.81	19.85	19.82
% SBP>160	8.52	8.55	8.58	8.54
% Treated	27.86	27.84	27.86	27.91
Mean Treatment Effect (SD)	-7.69 (3.8)	-15.33 (4.0)	-22.98 (4.2)	-30.65 (4.1)

Table 11: Descriptive statistics for Scenario 4.

1.2.5.2 Results

Figures 5 – 8 present the results for Scenario 4 graphically, while Appendix A contains the full tables of results in tables 37-40.

As in Scenario 3, the Substitution approaches [*Binary Trait* (d), *Fixed Substitution* (f), *Random Substitution* (g), and *Median Method* (h)] and *Exclude* (b) will not be affected in this scenario. These approaches either omit or substitute all observations corresponding to individuals on treatment and, hence, are uninfluenced by the changes to the treatment effect.

For the other approaches, the results in the sub-scenario with a 10% reduction in SBP due to treatment (see Figure 6) are also highly characteristic of those obtained in the General Simulation Study. The Informative Phenotype approaches [*Fixed Treatment Effect* (e), *Non-Parametric Adjustment* (i), and

Censored Normal Regression (j)] yield high powers (≈ 0.9) and only low levels of bias (e.g. mean $\hat{\beta}_3 \approx 1.7 - 2.1$), while results for the Naïve approaches [*No Adjustment* (a), *Exclude* (b), and *Treatment as a Binary Covariate* (c)] yield lower powers ($\approx 0.66 - 0.85$) and greater magnitudes of bias (e.g. mean $\hat{\beta}_3 \approx 1.25 - 1.55$). As stated earlier, the descriptive characteristics for this sub-scenario are very similar to those of the General Simulation Study. Hence, these results suggest that the way the treatment effect is modelled (i.e. as a proportional reduction of SBP or as a randomly generated reduction in SBP) makes little difference to how the methods perform.

In the other sub-scenarios, the performance of the approaches seems to depend on the magnitude of the treatment effect. When the treatment effect reduces SBP by 5% (see Figure 5), the approaches perform similarly – in terms of power – to the General Simulation Study, but in the sub-scenarios with treatment effects of 15% and 20% reductions of SBP (Figure 7 and Figure 8 respectively), the approaches generally yield lower powers than in the General Simulation Study. The level of bias also very much depends on the magnitude of the treatment effect. When the treatment effect is 5%, *No Adjustment* (a) actually yields only a small level of bias (mean $\hat{\beta}_3 = 1.78$), while the Informative BP approaches tend to overestimate the effect of g_1 [e.g. mean $\hat{\beta}_3 \approx 2.3$ for (i) and (j)]. Approach (a) clearly performs better in this sub-scenario, as the bias due to treatment use is small because the treatment effect itself is small. On the other hand, the Informative BP approaches tend to over-compensate for the use of treatment in this sub-scenario and, hence, end up overestimating the regression coefficient.

In the sub-scenarios with greater reductions of SBP due to treatment, all the approaches (other than the unaffected approaches listed above) tend to suffer from reduced powers and greater magnitudes of bias. This is simply because the treatment effects, on average, are greater in these scenarios and, hence, the bias due to the use of treatment is also greater. Although *Non-parametric Adjustment* (i) and *Censored Normal Regression* (j) still perform relatively well in these more extreme scenarios, *Fixed Treatment Effect* (e) suffers from an impaired performance. For example, when the treatment reduces SBP by 20% (mean treatment effect $\approx -30\text{mmHg}$) and when $c = 5$, (e) yields a reduced power of around 0.7 and drastically increased levels of bias (e.g. mean $\hat{\beta}_3 \approx 1.2$). Although previous scenarios have suggested that (e) is relatively insensitive to the value c , these results show that the approach does, in fact, require a sensible value for c . This conclusion, nonetheless, seems logical enough. For example, when $c = 5$ here, the imputed treatment effects are only around $1/6^{\text{th}}$ of the size of the true treatment effects. The corrections to *modified BPs* imposed by this approach are therefore insufficient here. This finding indicates that any use of approach (e) in practice should base the choice of the value c on external and expert knowledge in order to ensure that a suitable value be used.

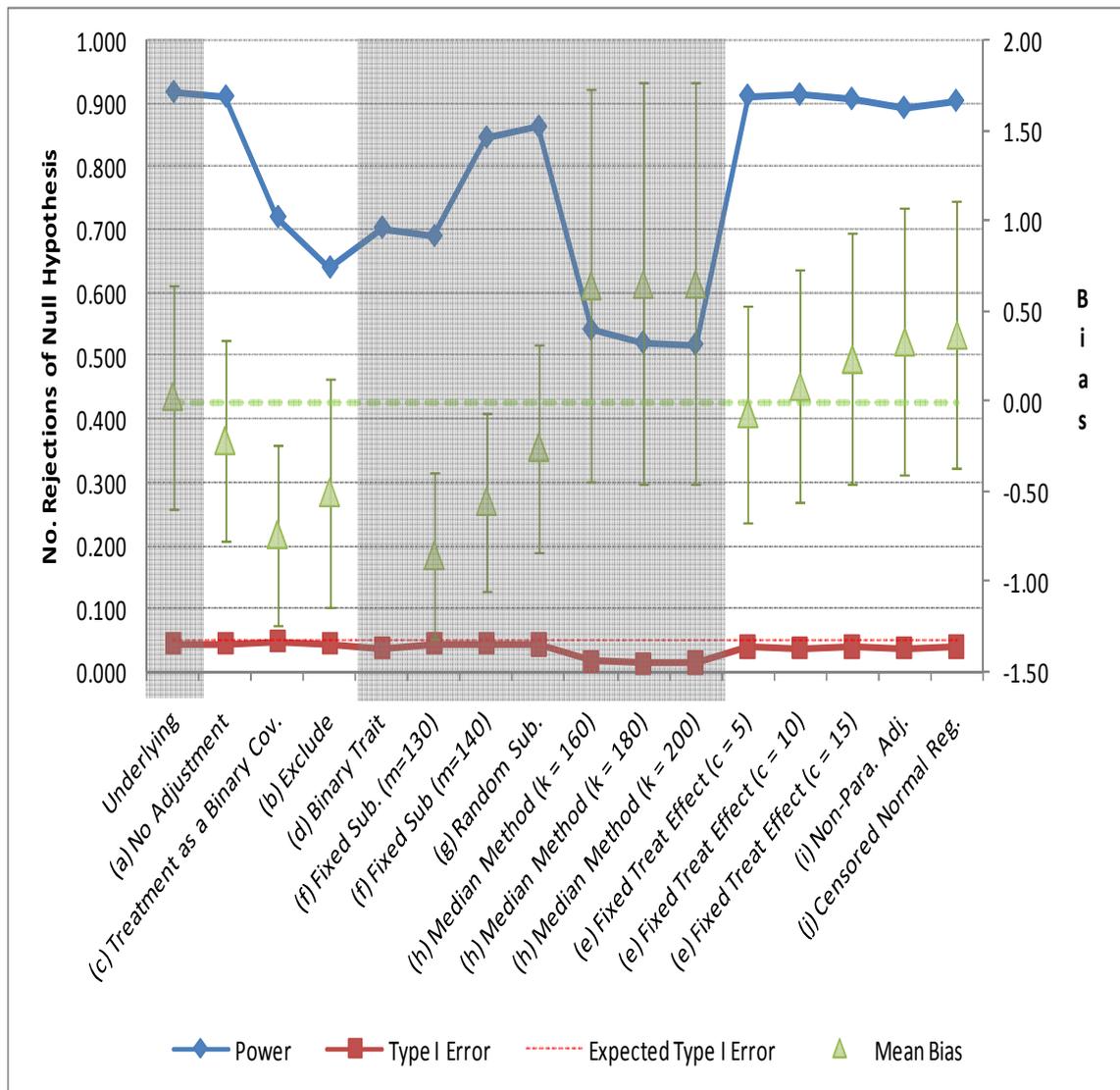


Figure 5: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 5%. Approaches are arranged in the categories (from left to right): naïve, substitution, informative phenotype. Power (relative to the gene parameter, β_3) and Type I Error (relative to the gene2 parameter, β_4) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

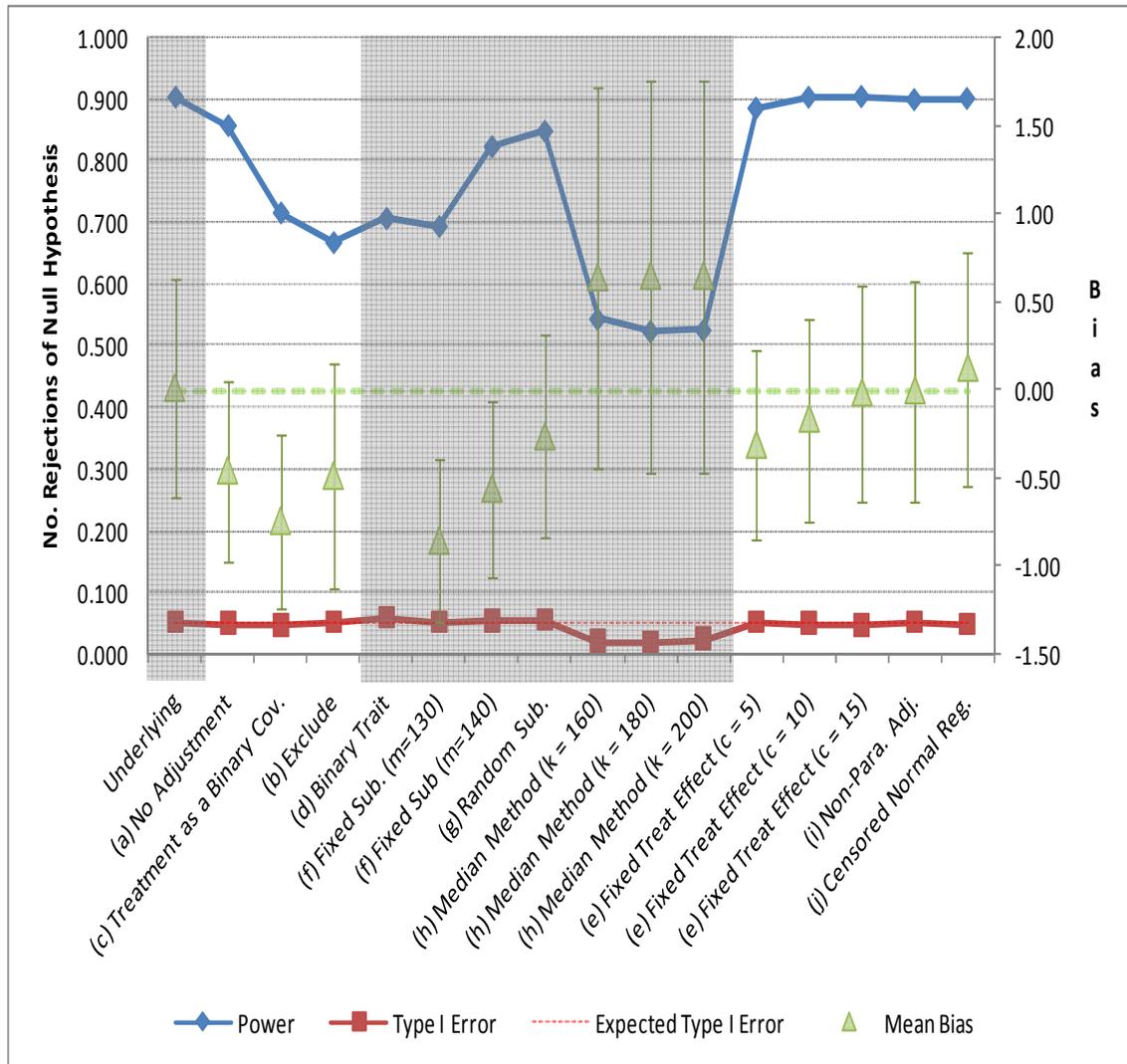


Figure 6: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 10%. Approaches are arranged in the categories (from left to right): naïve, substitution, informative phenotype. Power (relative to the gene parameter, β_3) and Type I Error (relative to the gene2 parameter, β_4) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

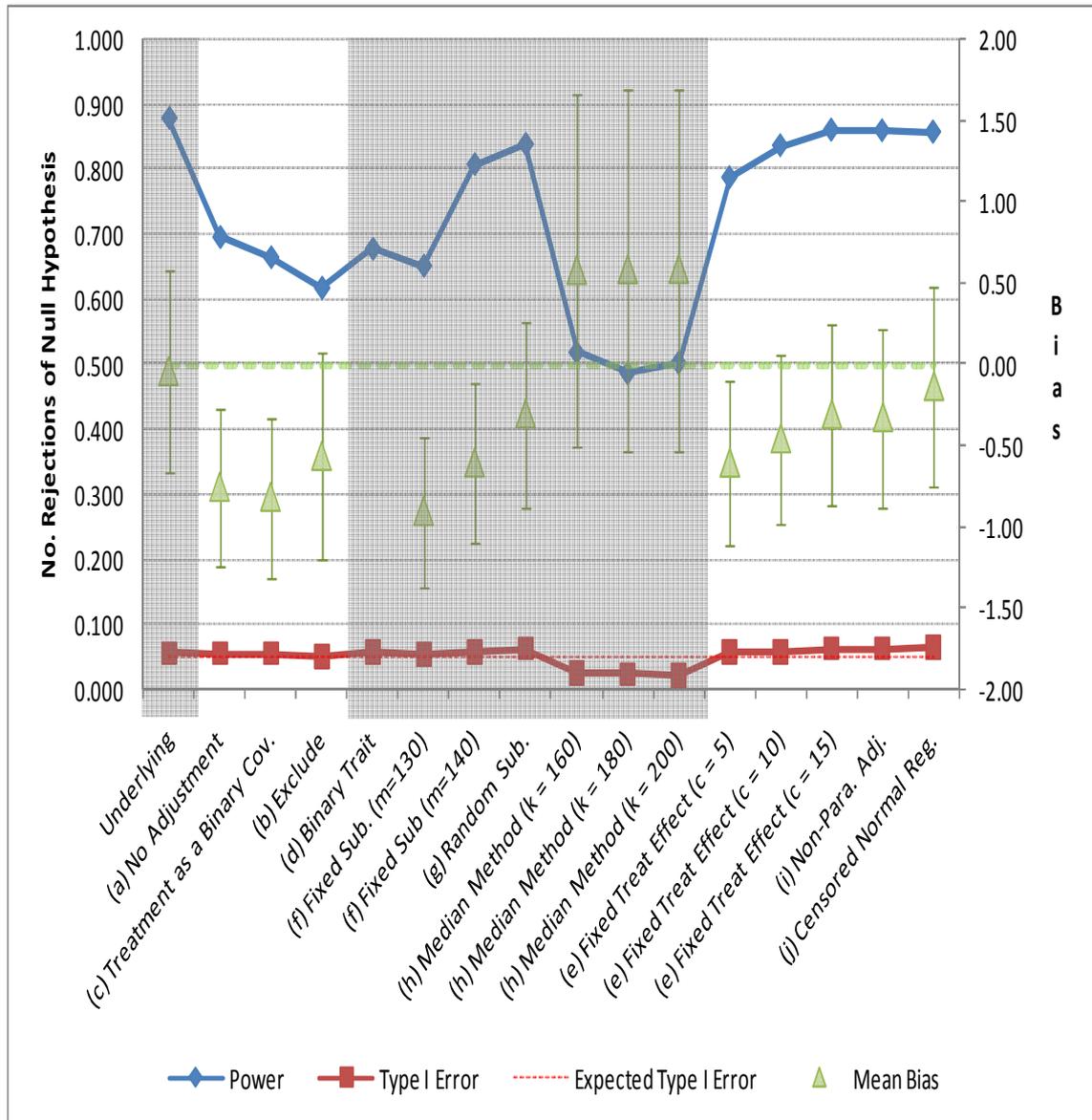


Figure 7: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 15%. Approaches are arranged in the categories (from left to right): naïve, substitution, informative phenotype. Power (relative to the gene parameter, β_3) and Type I Error (relative to the gene2 parameter, β_4) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

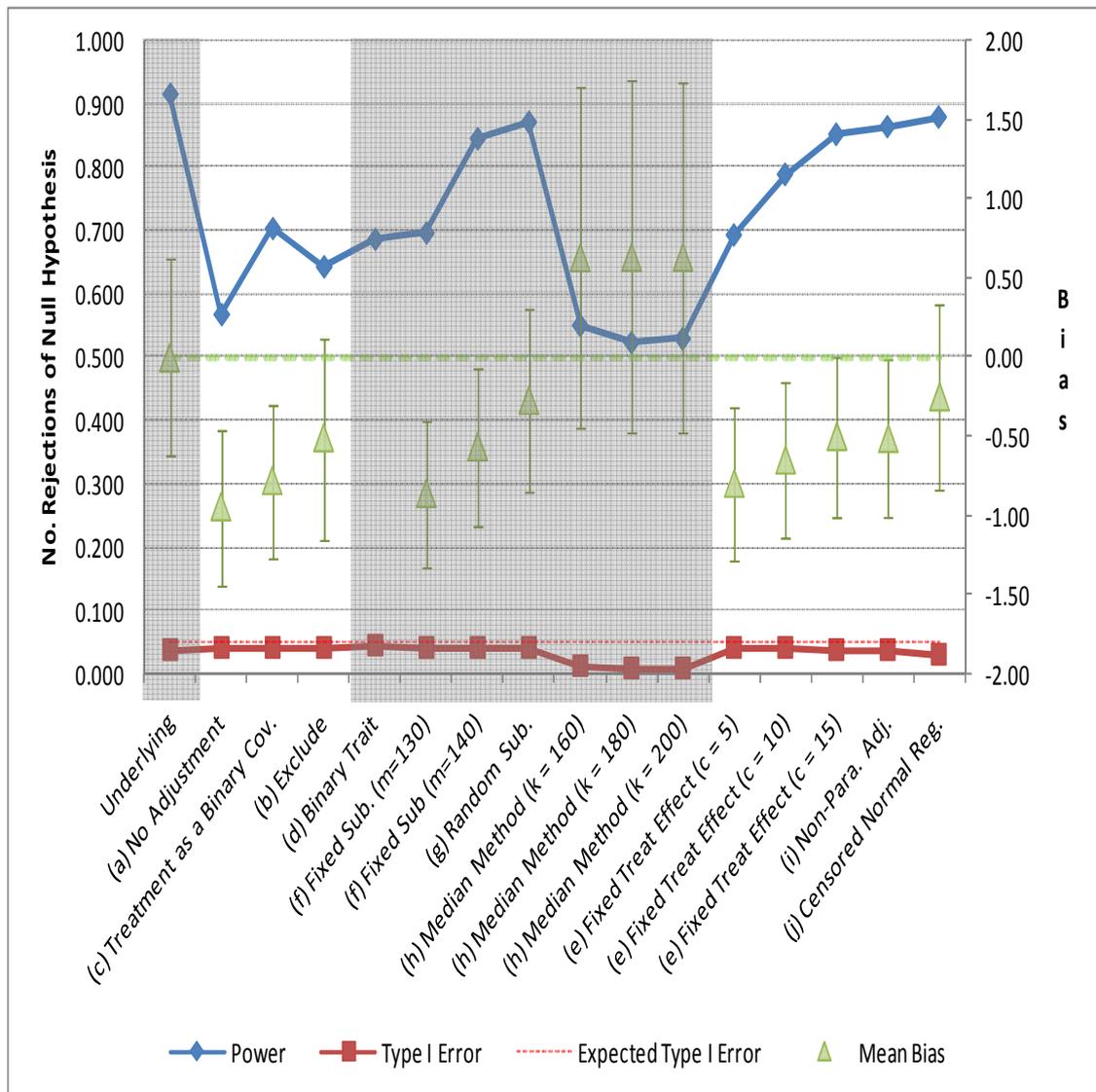


Figure 8: Graphical representation of results for Scenario 4 with a proportional reduction of SBP of 20%. Approaches are arranged in the categories (from left to right): naïve, substitution, informative phenotype. Power (relative to the gene parameter, β_3) and Type I Error (relative to the gene2 parameter, β_4) are evaluated on the left vertical axis, and mean bias/SE (with respect to β_3) is evaluated on the right vertical axis.

1.3. Simulation Studies: Differential intervention

The previous section tests a number of scenarios under a non-differential intervention, i.e. where both the probability of receiving treatment and the effect of treatment depend only upon BP. In contrast, this section focuses on testing how the approaches to analysis perform when the intervention is *differential*.

Scenarios 5 and 6 investigate the effects of differential treatment effects, while Scenario 7 investigates the effects of a differential probability for receiving treatment.

1.3.1 Scenario 5: Pharmacogenetic Interaction

This scenario aims to determine how the approaches perform when the effect of treatment depends on another factor other than BP. This situation would occur, for example, if any interaction exists between a genetic variant and a treatment. Interactions between genes and treatments are known as *pharmacogenetic interactions*, and there is already evidence of the existence of pharmacogenetic interactions with antihypertensive treatments (Wang *et al.*, 2007b). Although the term “pharmacogenetic interaction” can imply an experience of differential side-effects to treatment or a differential response to treatment by genotype, only the latter case is relevant here. This scenario therefore tests the approaches under a condition where a pharmacogenetic interaction influences the efficacy of treatment.

1.3.1.1 Simulation Method

Pharmacogenetic interactions that influence the magnitude of BP lowering due to treatment can occur in one of two possible directions: those with the minor allele for a particular genetic variant could be more sensitive or less sensitive to the treatment. This scenario therefore simulates two sub-scenarios demonstrating the effects of a pharmacogenetic interaction in each direction. These sub-scenarios are referred to as scenarios 5a and 5b.

To remain consistent with earlier scenarios, this scenario broadly uses the same simulation method as the General Simulation Study (see Section 1.2.1). However, some properties must be altered to allow for the inclusion of a pharmacogenetic interaction. For instance, the aim of these analyses is to evaluate the approaches in terms of their capabilities to detect and estimate the marginal effect of a single nucleotide polymorphism (SNP) on SBP – when that SNP interacts with treatment. As such, both type I error and power must be evaluated with respect to this SNP. Hence, in contrast to previous scenarios, which simulate two genetic variants to evaluate power and type I error respectively, this scenario simulates a single SNP only. Power is evaluated by running the simulation 1,000 times with a SNP effect of +2 mmHg, and type I error is evaluated by running the simulation a further 1,000 times with a genetic effect of 0. In addition, a third SNP effect of -2 mmHg is also tested in this scenario, in order to demonstrate the relationship between the sign of the SNP effect, the direction of the pharmacogenetic interaction, and the statistical power. Therefore, the simulation is also performed 1,000 times with a simulated SNP effect of -2 mmHg.

This scenario generates underlying SBP from a linear regression model similar to Equation 12, with the only difference being that a single genetic variant, g , is now included in the model instead of two variants (see Equation 13 below). Here Z_i is the underlying SBP for the i 'th individual ($i = 1, \dots, 2000$); AGE_i denotes age (25-80 years); SEX_i denotes sex (1=male, 0=female); g_i denotes genotype for a diallelic locus with an allele frequency of 0.3 (0, 1 or 2 copies of the minor allele); and ε_i denotes random error, the simulation model is therefore:

Equation 13

$$Z_i = \beta_0 + \beta_1 AGE_i + \beta_2 SEX_i + \beta_3 g_i + \varepsilon_i,$$

where $\beta_1 = 0.4$, $\beta_2 = 3$, and $\varepsilon_i \sim N(0,18)$. Because the SNP effect, β_3 , is either +2, -2 or 0, the constant term, β_0 , is adjusted accordingly to maintain a consistent distribution for Z_i as in the General Simulation Study. Hence, $\beta_0 = 110, 112.4$, and 111.2 for $\beta_3 = +2, -2$, and 0 respectively.

As before, only hypertensives (i.e. those with underlying SBP ≥ 140 mmHg) receive treatment here. Treatment is again allocated to these individuals with a probability of 0.75, while *normotensives* (i.e. those with underlying SBP < 140 mmHg) receive treatment with a probability of 0.

At this point, the pharmacogenetic interaction is introduced by varying the treatment effect by genotype for g . In contrast to the General Simulation Study, which generates the treatment effect randomly for each individual from $N(-15, 4^2)$, this scenario generates the treatment effect from one of three distributions that correspond to each genotype for g . For instance, an individual with no copies of the minor allele for g is drawn a treatment effect from distribution one; an individual with one copy of the minor allele is drawn a treatment effect from distribution two; and an individual with two copies of the minor allele is drawn a treatment effect from distribution three. The magnitude of the treatment effect, thus, now depends on a subject's genotype for g .

To ensure consistency between the distributions of observed SBP generated in this scenario and the General Simulation Study, the treatment effects generated here have a mean of -15 mmHg over all individuals, i.e. the treatment effects

are centred at -15 mmHg. Furthermore, each of the three distributions for the treatment effect used here have the same variance as the treatment effect simulated in the General Simulation Study (i.e. $\text{var} = 4^2$). Scenario 5a simulates a pharmacogenetic interaction that *reduces* the efficacy of treatment in the presence of the minor allele, while Scenario 5b simulates a pharmacogenetic interaction that *increases* the efficacy of treatment in the presence of the minor allele. Hence, in Scenario 5a the treatment effect, γ_i , is generated from $N(18, 4^2)$, $N(13.43, 4^2)$ or $N(9, 4^2)$ corresponding to whether the i^{th} individual has 0, 1, or 2 copies of the minor allele respectively; and in Scenario 5b, γ_i is generated from $N(12, 4^2)$, $N(17.43, 4^2)$ or $N(20, 4^2)$ respectively.

Finally, the observed SBP, Y_i , is derived – as in the General Simulation Study – by subtracting the treatment effect, γ_i , from Z_i where applicable.

Table 12 below lists the full simulation properties for Scenario 5.

Parameter	Scenario 5a	Scenario 5b	
Simulation Runs	1000	1000	
Sample Size	2000	2000	
Constant	110/112.4/111.2	110/112.4/111.2	
Age (years): age_i ; β_1	Uniform [25-80]; 0.4	Uniform [25-80]; 0.4	
Sex (Male/Female): sex_i ; β_2	Bernoulli (0.5); 3	Bernoulli (0.5); 3	
Gene: g_i ; β_3	Bin(2, 0.3); +2/-2/0	Bin(2, 0.3); +2/-2/0	
Random Error	$\sim N(0, 18)$	$\sim N(0, 18)$	
Hypertension Criterion	$SBP \geq 140$	$SBP \geq 140$	
$P\{treat_i=1 \mid \text{Hypertensive}\}$	0.75	0.75	
Mean Treatment Effect (mmHg)	15	15	
Pharmacogenetic Interaction (Treatment effect)	If 0 copies minor-allele	$\sim N(18, 4^2)$	$\sim N(12, 4^2)$
	If 1 copy minor-allele	$\sim N(13.43, 4^2)$	$\sim N(17.43, 4^2)$
	If 2 copies minor-allele	$\sim N(9, 4^2)$	$\sim N(20, 4^2)$

Table 12: Simulation properties for Scenario 5.

Note that as the treatment effects have been centred here, scenarios 5a and 5b yield similar descriptive statistics to one another. Table 13 below therefore presents descriptive statistics for Scenario 5a only, based on 1,000 simulation runs. Note also that due to the adjustment of the intercept parameter to account for the varying effect of the genetic variant, g , the descriptive statistics are stable for each of the three simulated effect sizes, and directly compare to the descriptive statistics presented for the General Simulation Study in Table 4.

Summary Statistic		Scenario 5a		
		Gene = 2	Gene = -2	Gene = 0
Mean Underlying SBP (SD) (mm Hg)	Overall:	133.68 (19.2)	133.70 (19.2)	133.70 (19.1)
	[$Treat_i = 1$]:	153.23 (10.4)	153.27 (10.5)	153.19 (10.4)
	[$Treat_i = 0$]:	126.13 (16.3)	126.15 (16.2)	126.17 (16.2)
Mean Observed SBP (SD) (mm Hg)	Overall:	129.48 (16.1)	129.39 (16.0)	129.45 (15.9)
	[$Treat_i = 1$]:	138.15 (11.7)	137.79 (11.5)	137.92 (11.5)
	[$Treat_i = 0$]:	126.13 (16.3)	126.15 (16.2)	126.17 (16.2)
% SBP>140		37.17	37.16	37.14
% SBP>150		19.81	19.84	19.75
% SBP>160		8.51	8.55	8.49
% Treated		27.87	27.85	27.86
Mean Treatment Effect (SD) (mm Hg)		15.08 (5.0)	15.48 (4.9)	15.27 (4.9)

Table 13: Descriptive statistics for Scenario 5a (based on 1,000 simulation runs).

1.3.1.2 Results

As with the analyses reported in Section 1.2, the focus of these analyses is on estimating and detecting the marginal effect of the genetic variant g on BP. Each approach fits the model in Equation 13 to the estimates of the underlying SBP, X_i [with the exception of *Treatment as a Binary Covariate* (c) and *Binary Trait* (d) – see Section 1.2.1.2 for an explanation]. For a recap of the approaches, see Table 3 (Page 32).

Results for scenarios 5a and 5b are summarised graphically in figures 9 and 10 respectively. As before, each figure shows the statistical power to detect the marginal effect of g and the type I error for each approach on the left-vertical axis (at the 5% level of significance). The mean bias of the estimated coefficient of g (in mmHg), with standard error, is shown for each approach on

the right-vertical axis. The figures show the statistical power relative to the simulated g effects of +2 mm Hg and -2 mm Hg per copy of the minor allele (i.e. $\beta_3 = 2$ and $\beta_3 = -2$ respectively), but the mean bias is displayed relative to $\beta_3 = 2$ only. Results are based on 1,000 runs for each scenario. The full tables of results are presented in tables 41-43 of Appendix A for Scenario 5a, and in tables 44-46 in Appendix A for Scenario 5b (for $\beta_3 = 2, -2,$ and 0 respectively).

The approaches to analysis are arranged across the x-axis in group order: Naïve [*No Adjustment* (a), *Exclude* (b), and *Treatment as a Covariate* (c)], Substitution [*Binary Trait* (d), *Fixed Substitution* (f), Random Substitution (g), and the *Median Method* (h)], and Informative BP [*Fixed Treatment Effect* (e), *Non-parametric Adjustment* (i), *Censored Normal Regression* (j)]. As with the figures in Section 1.2, the additional analysis of underlying SBP is also shown, for comparison purposes, on the far-left of the x-axis. This scenario also reports results for a further additional analysis, which extends approach (c). This additional analysis again models treatment as a binary covariate, but, in addition, it also explicitly includes the SNP-treatment interaction term. Results for this additional analysis are therefore represented in figures 9 and 10 adjacent to the results for (c).

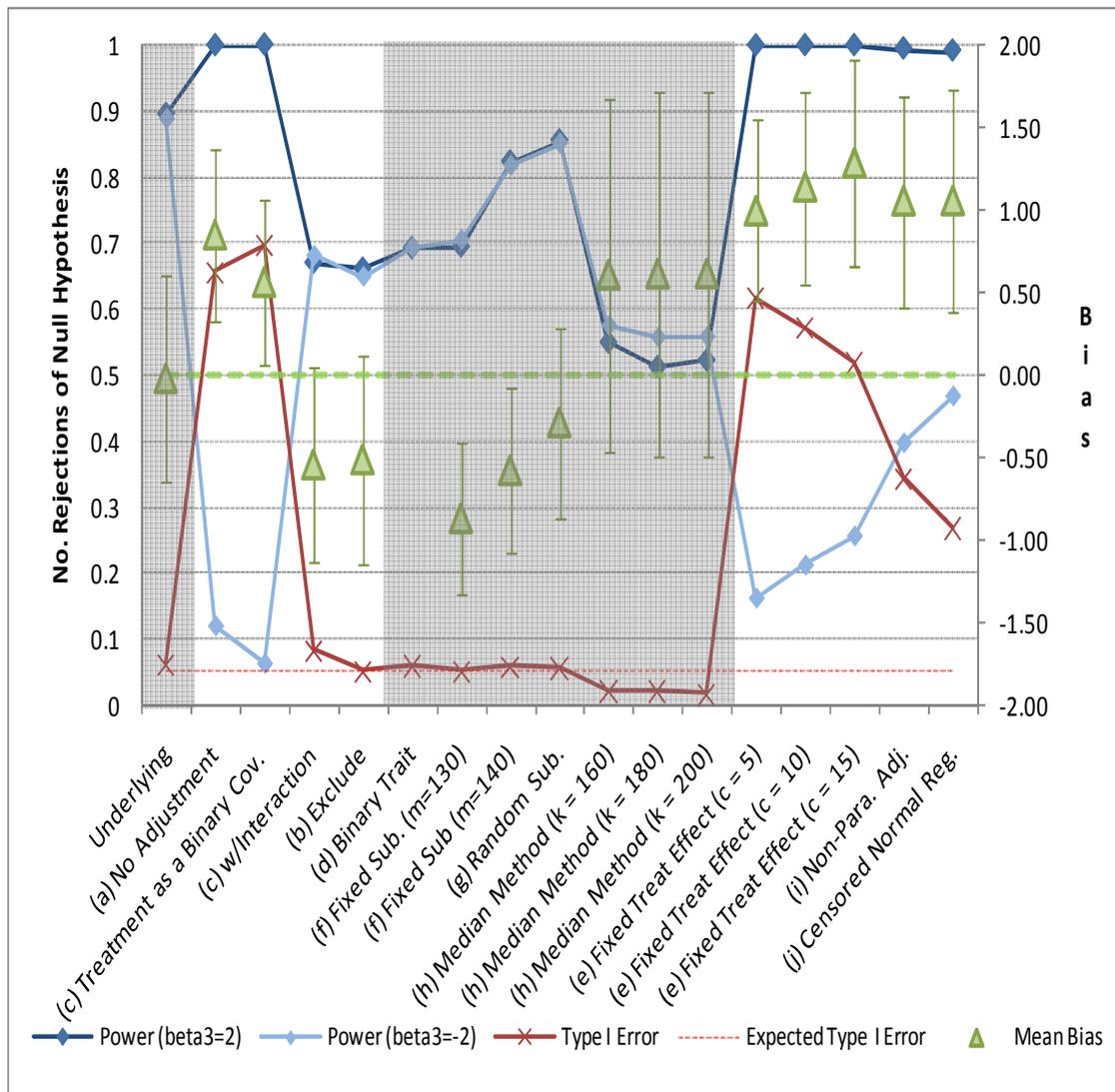


Figure 9: Graphical representation of the results for Scenario 5a. Approaches are arranged here in categories (from left to right): naive, substitution, informative phenotype. Power to detect the genetic variant, g , when $\beta_3 = 2$ and $\beta_3 = -2$ are denoted in dark blue and light blue diamonds respectively, and type I error relative to the genetic effect is denoted in red crosses. Power and type I error are evaluated on the left vertical axis. Mean bias/SE with respect to $\beta_3 = 2$ is shown in green triangles, and is evaluated on the right vertical axis.

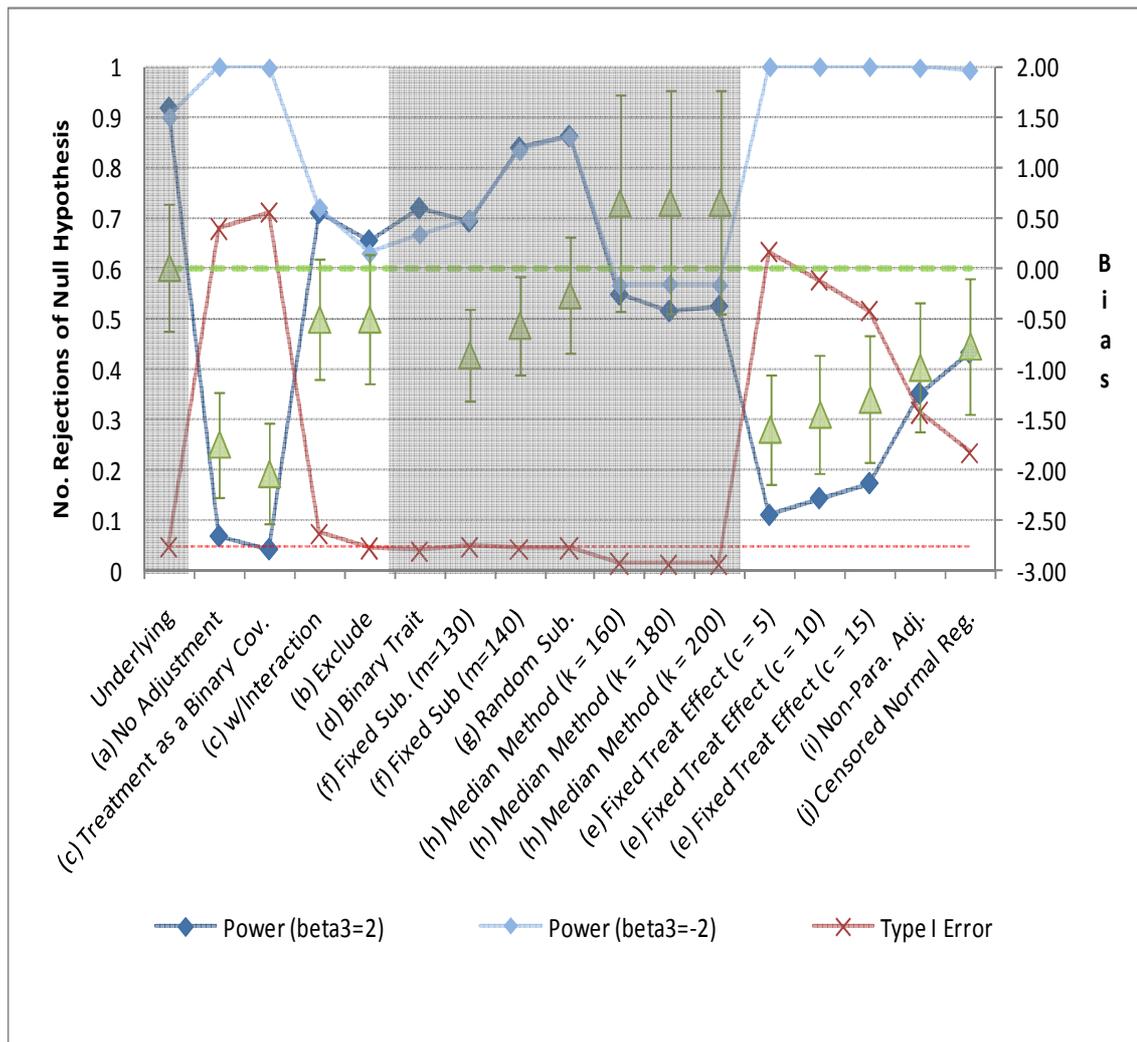


Figure 10: Graphical representation of the results for Scenario 5b. Approaches are arranged here in categories (from left to right): naïve, substitution, informative phenotype. Power to detect the genetic variant, g , when $\beta_3 = 2$ and $\beta_3 = -2$ are denoted in dark blue and light blue diamonds respectively, and type I error relative to the genetic effect is denoted in red crosses. Power and type I error are evaluated on the left vertical axis. Mean bias/SE with respect to $\beta_3 = 2$ is shown in green triangles, and is evaluated on the right vertical axis.

Inspection of the results shown in Figure 9 and Figure 10 reveals a clear divide between how the different classes of approaches perform when there is a differential response to treatment. As in scenarios 3 and 4, the Substitution approaches and *Exclude* (b) all yield comparable results to the General Simulation Study here. These approaches avoid the effects of the differential treatment effect because they replace *modified BPs* with alternative values

derived independently from the observed data [or, in the case of (b), it excludes any *modified BPs* from the analysis]. Hence, these approaches are unaffected by the pharmacogenetic interaction simulated with g .

In contrast, the effects of the pharmacogenetic interaction upon the Informative BP approaches and *No Adjustment* (a) and *Treatment as a Binary Covariate* (c) are striking. For instance, in Figure 9 each of these approaches markedly overestimates the marginal effect of g and in Figure 10 each of these approaches markedly *underestimates* g [mean bias ≈ 0.5 to 2 mmHg for (a) and (c); mean bias ≈ 0.5 to 1.5 mmHg for (e), (i) and (j)]. Consequently, the power and type I error rates for these approaches are also affected. In Figure 9, when the simulated effect of the genetic variant g is 2 mmHg ($\beta_3 = 2$), the power for each of these approaches is increased ($\approx 100\%$), but when $\beta_3 = -2$, the powers are reduced substantially [power $\approx 10\%$ for (a) and (c); power $\approx 18\% - 48\%$ for (e), (i) and (j)]. The type I error rates for these approaches are also similarly affected. In contrast to the General Simulation Study – where each approach yields the correct level of type I error, the type I error rates here are highly elevated [type I error ≈ 0.7 for (a) and (b); type I error $\approx 0.3 - 0.6$ for (e), (i) and (j)].

Given that these approaches do not account for what is effectively a SNP-treatment interaction, these patterns of results are not entirely surprising. For example, in Scenario 5a, a reduced treatment effect is associated with the minor allele at g . Hence, ignoring the true main effect of g , individuals homozygous for (i.e. possessing two copies of) the minor allele who receive antihypertensive treatment will be less responsive to the treatment and will, on

average, have greater *modified BP* than treated individuals who are homozygous for the *major* allele. In contrast, Scenario 5b simulates a pharmacogenetic interaction that *increases* the efficacy of treatment in presence of the minor allele. Thus, again ignoring the true effect of g , treated individuals homozygous for the minor allele will, on average, have *lower* modified BP than treated individuals without the minor allele. Consequently, estimates of the marginal effect of g are biased upwards in Scenario 5a and biased downwards in Scenario 5b.

There is a clear relationship here between the power of these affected approaches, the sign of the simulated g effect, and the direction of the pharmacogenetic interaction. For instance, these approaches can yield either an increased or a decreased power in these settings, depending on whether the sign of the main effect of g agrees with or contradicts the direction of the interaction. As described above, if a pharmacogenetic interaction reduces the efficacy of treatment, the marginal effect of g will always be overestimated. Hence, if the main effect of g is positive the power will be increased, but if the main effect of g is negative the power will be decreased. If, on the other hand, the pharmacogenetic interaction *increases* the efficacy of treatment in presence of the minor allele, the marginal effect of g will be underestimated. Thus, the opposite pattern of results to the above applies.

Results for the additional analysis performed in this scenario, which models the SNP-treatment interaction term in addition to the treatment main effect, yields similar results here to *Treatment as a Binary Covariate* (c) in the General Simulation Study. This approach is the only approach that utilises the modified

BPs and yet remains unaffected by the pharmacogenetic interaction. Nevertheless, despite yielding approximately the correct level of type I error here, a low statistical power ($\approx 70\%$) is obtained and its estimates of the regression coefficients are, on average, shrunken to the null. This approach avoids the problems of some of the other approaches in this scenario because it explicitly models the SNP-treatment interaction term. This is actually the only approach that can account for potential SNP-treatment interactions in this way, because, in order to do this, the treatment main-effect must also be fitted. For the reasons discussed in Section 1.1.2, adjusting for treatment by modelling the treatment term as a binary covariate is a flawed approach to the analysis of BP. Hence, this approach, too, is a suboptimal approach to analysis.

In addition to the results described above, a further feature of the full tables of results shown in Appendix A concerns the estimates of the regression coefficients for the parameters not involved in the pharmacogenetic interactions, i.e. AGE and SEX. For the Informative BP approaches [i.e. (e), (i) and (j)], which perform well in previous scenarios but yield biased estimates of the main effect of the genetic variant, g , here, estimates of the effects of the other regression coefficients are unperturbed. This observation, hence, demonstrates that only estimates of the effects of parameters involved in interactions with treatment are biased in these situations. The implication of this is that, even if a pharmacogenetic interaction exists, the Informative BP approaches should provide reasonable estimates of any genetic variants or other explanatory variables of interest that are not involved in the interaction. An additional scenario was simulated to confirm these observations (results not provided), in which estimates of the power and type I error were derived with respect to an

alternative genetic variant to the one involved in the pharmacogenetic interaction. The Informative BP approaches (as well as every other approach) produced similar results in this additional scenario to those obtained in the General Simulation Study. Hence, these results support the above conclusions.

1.3.2 Scenario 6: Pharmacogenetic Interaction with One Class of Treatment

The pharmacogenetic interaction implemented in the previous scenario implicates all subjects on treatment. Given that different classes of antihypertensive medication exist, and that these act upon different biological pathways, it is probably unrealistic to assume that a given genetic variant will interact with all treatment types. Hence, because, in reality, different subjects will use different classes of treatment, the influence of the pharmacogenetic interaction simulated in Scenario 5 is likely to be more extreme than that of a real pharmacogenetic interaction in a real genetic association study of BP. This scenario therefore simulates two different classes of treatment, of which only one interacts with the genetic variant, g .

1.3.2.1 Simulation Method

The simulation method used for this scenario is based on that used in Scenario 5. As before, the underlying SBP, Z_i , is generated from the model in Equation 13, and hypertensive individuals are allocated treatment with probability 0.75. In this scenario, however, two classes of treatment are simulated. Participants selected to receive treatment are subsequently randomised either to receive Treatment A or Treatment B, with the probabilities 0.33 and 0.67 respectively.

Treatment A represents angiotensin-converting enzyme (ACE) inhibitors, a common class of antihypertensive medication (Wang *et al.*, 2007b), and Treatment B represents usage of any other antihypertensives (pooled together). For Treatment A, the treatment effect depends on the genetic variant, g , and a pharmacogenetic interaction is implemented in the same way as that implemented in Scenario 2. Hence, for individuals on Treatment A, γ_i , is generated from $N(18, 4^2)$, $N(13.43, 4^2)$, or $N(9, 4^2)$ corresponding to whether the i^{th} individual has 0, 1, or 2 copies of the minor allele respectively; for individuals on Treatment B, the treatment effect is independent of g , and is thus generated as in the General Simulation Study [i.e. $\gamma_i \sim N(15, 4^2)$]. All treatment effects are again truncated at zero.

As in Scenario 5, a pharmacogenetic interaction is simulated in both directions, and effect sizes of +2, -2 and 0 are tested for the genetic variant, g . Scenario 6a refers to the situation where the pharmacogenetic interaction reduces the magnitude of SBP lowering due to Treatment A, and Scenario 6b refers to the situation where the pharmacogenetic interaction increases the efficacy of Treatment A. As before, the additional analysis that extends *Treatment as a Binary Covariate* (c) by modelling the SNP-treatment interaction term is also performed here.

Table 14 lists the simulation properties for Scenario 6.

Parameter	Scenario 6a	Scenario 6b	
Simulation Runs	1000	1000	
Sample Size	2000	2000	
Constant	110/112.4/111.2	110/112.4/111.2	
Age (years): age_i ; β_1	Uniform [25-80]; 0.4	Uniform [25-80]; 0.4	
Sex (Male/Female): sex_i ; β_{12}	Bernoulli (0.5); 3	Bernoulli (0.5); 3	
Gene: g_i ; β_{13}	Bin(2, 0.3); +2/-2/0	Bin(2, 0.3); +2/-2/0	
Random Error	$\sim N(0, 18)$	$\sim N(0, 18)$	
Hypertension Criterion	SBP \geq 140	SBP \geq 140	
P { $treat_i=1$ Hypertensive}	0.75	0.75	
P {Treatment A $treat_i=1$ }	0.33	0.33	
P{Treatment B $treat_i=1$ }	0.67	0.67	
Mean Effect Treatment A (mmHg)	15	15	
Pharmacogenetic Interaction (Treatment A Effect) [mmHg]	If 0 copies minor-allele	$\sim N(18, 4^2)$	$\sim N(12, 4^2)$
	If 1 copy minor-allele	$\sim N(13.43, 4^2)$	$\sim N(17.43, 4^2)$
	If 2 copies minor-allele	$\sim N(9, 4^2)$	$\sim N(20, 4^2)$
Treatment B Effect (mmHg)	$\sim N(15, 4^2)$	$\sim N(15, 4^2)$	

Table 14: Simulation Properties for Scenario 6.

The descriptive statistics for Scenario 6a are summarised in Table 15 below. As the treatment effects are again centred in this scenario, Scenario 6b yields similar descriptive statistics and, hence, these are not provided.

Summary Statistic		Scenario 6a		
		Gene = 2	Gene = 0	Gene = - 2
Mean Underlying SBP (SD) (mmHg)	Overall:	133.72 (19.2)	133.67 (19.2)	133.70 (19.1)
	Subjects on Treatment A:	153.24 (10.4)	153.27 (10.4)	153.20 (10.4)
	Subjects on Treatment B:	153.26 (10.4)	153.24 (10.5)	153.24 (10.4)
	<i>Untreated</i> subjects:	131.71 (18.8)	131.66 (18.8)	131.71 (18.7)
Mean Observed SBP (SD) (mmHg)	Overall:	129.52 (16.0)	129.46 (16.0)	129.50 (15.9)
	Subjects on Treatment A:	138.19 (11.7)	137.81 (11.4)	137.93 (11.5)
	Subjects on Treatment B:	138.25 (11.2)	138.25 (11.2)	138.24 (11.2)
	<i>Untreated</i> subjects:	128.63 (16.1)	128.60 (16.1)	128.64 (16.1)
% SBP>140		37.17	37.11	37.10
% SBP>150		19.87	19.81	19.77
% SBP>160		8.57	8.55	8.52
% Treatment A		9.33	9.30	9.27
% Treatment B		18.58	18.52	18.56
Mean Treatment A Effect (SD) (mmHg)		15.06 (5.0)	15.46 (4.9)	15.27 (4.95)
Mean Treatment B Effect (SD)(mmHg)		15.01 (4.0)	14.99 (4.0)	15.00 (4.0)

Table 15: Descriptive statistics for Scenario 6a.

1.3.2.2 Results

Figures 11 and 12 present the results in graphical form for scenarios 6a and 6b respectively, while the full tables of results are presented in Appendix A (tables 47-49 for Scenario 6a; tables 50-52 for Scenario 6b).

As discussed above, the pharmacogenetic interactions implemented in this scenario affect only the participants on Treatment A, and, thus, have more moderate effects than the extreme pharmacogenetic interactions simulated in Scenario 5. For instance, *No Adjustment (a)*, *Treatment as a Binary Covariate (c)*, and the Informative BP approaches [*Fixed Treatment Effect (e)*, *Non-Parametric Adjustment (i)* and *Censored Normal Regression (j)*] generally yield

less bias in this scenario than in Scenario 5, and, consequently, the statistical powers and type error rates are less badly affected. Nevertheless, the type I error rates for these approaches remain above 5% ($\approx 8\% - 13\%$), and, when the direction of the pharmacogenetic interaction conflicts with the sign of the effect of the genetic variant, g , the statistical powers remain substantially lower than those obtained in the General Simulation Study ($\approx 40\% - 80\%$).

As in the previous scenario, the Substitution approaches [*Binary Trait* (d), *Fixed Substitution* (f), *Random Substitution* (g) and *Median Method* (h)] and *Exclude* (b) remain completely unaffected by the pharmacogenetic interaction here, and perform the same as in the General Simulation Study. The additional analysis that models the SNP-treatment interaction is also unaffected, and again yields similar results here to (c) in the General Simulation Study.

This scenario is fundamentally based on the same simulation method as in Scenario 5 and, as such, the results can be explained in the same way. The Informative BP approaches and (a) and (c) are affected here because they utilise all the observed data and apply the same correction for treatment (where applicable) to all participants, regardless of their genotype or treatment class. Because the genetic variant of interest, g , interacts with Treatment A, estimates of the effect of g on BP – over both treated and non-treated subjects – are biased away from its true (i.e. marginal) effect.

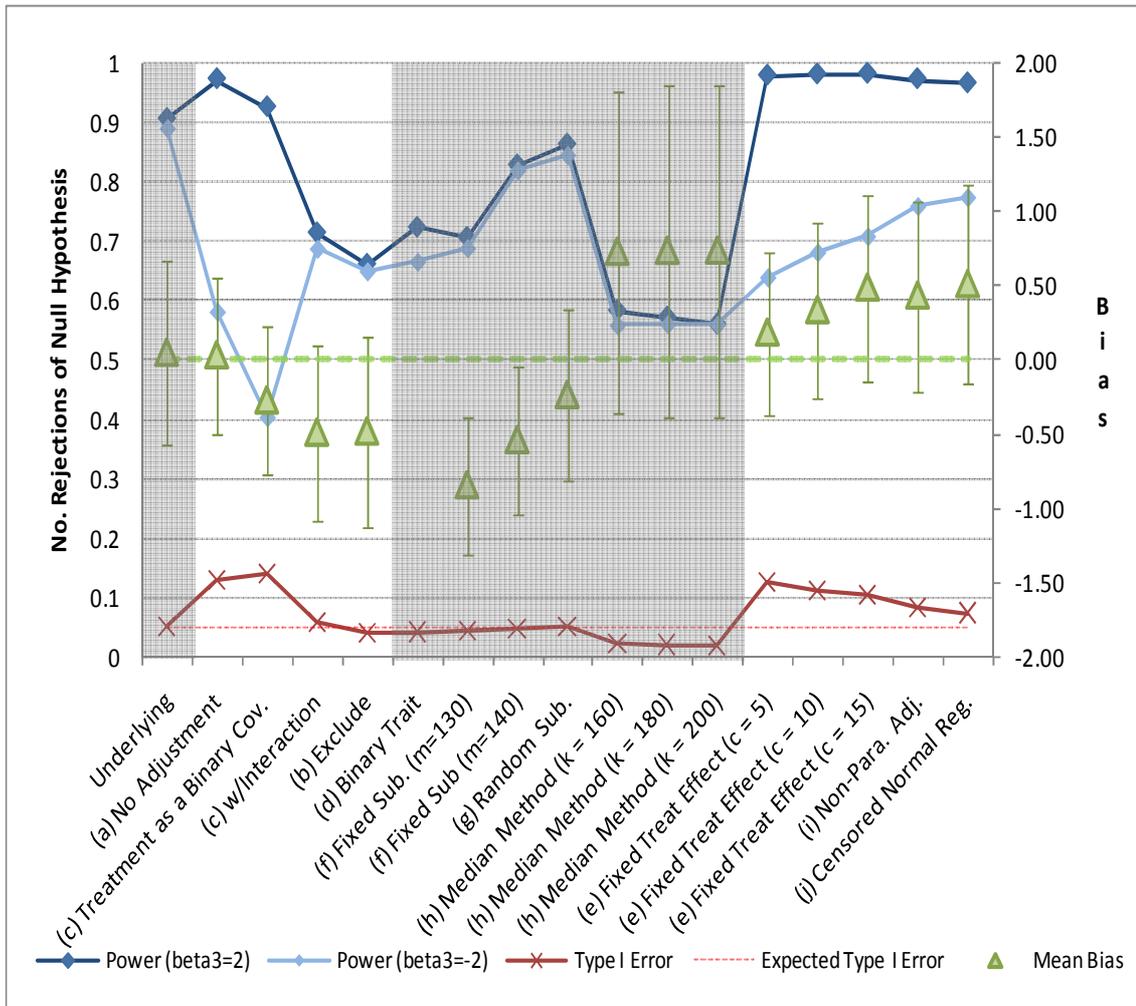


Figure 11: Graphical representation of the results for Scenario 6a. Approaches are arranged here in categories (from left to right): naïve, substitution, informative phenotype. Power to detect the genetic variant, g , when $\beta_3 = 2$ and $\beta_3 = -2$ are denoted in dark blue and light blue diamonds respectively, and type I error relative to the genetic effect is denoted in red crosses. Power and type I error are evaluated on the left vertical axis. Mean bias/SE with respect to $\beta_3 = 2$ is shown in green triangles, and is evaluated on the right vertical axis.

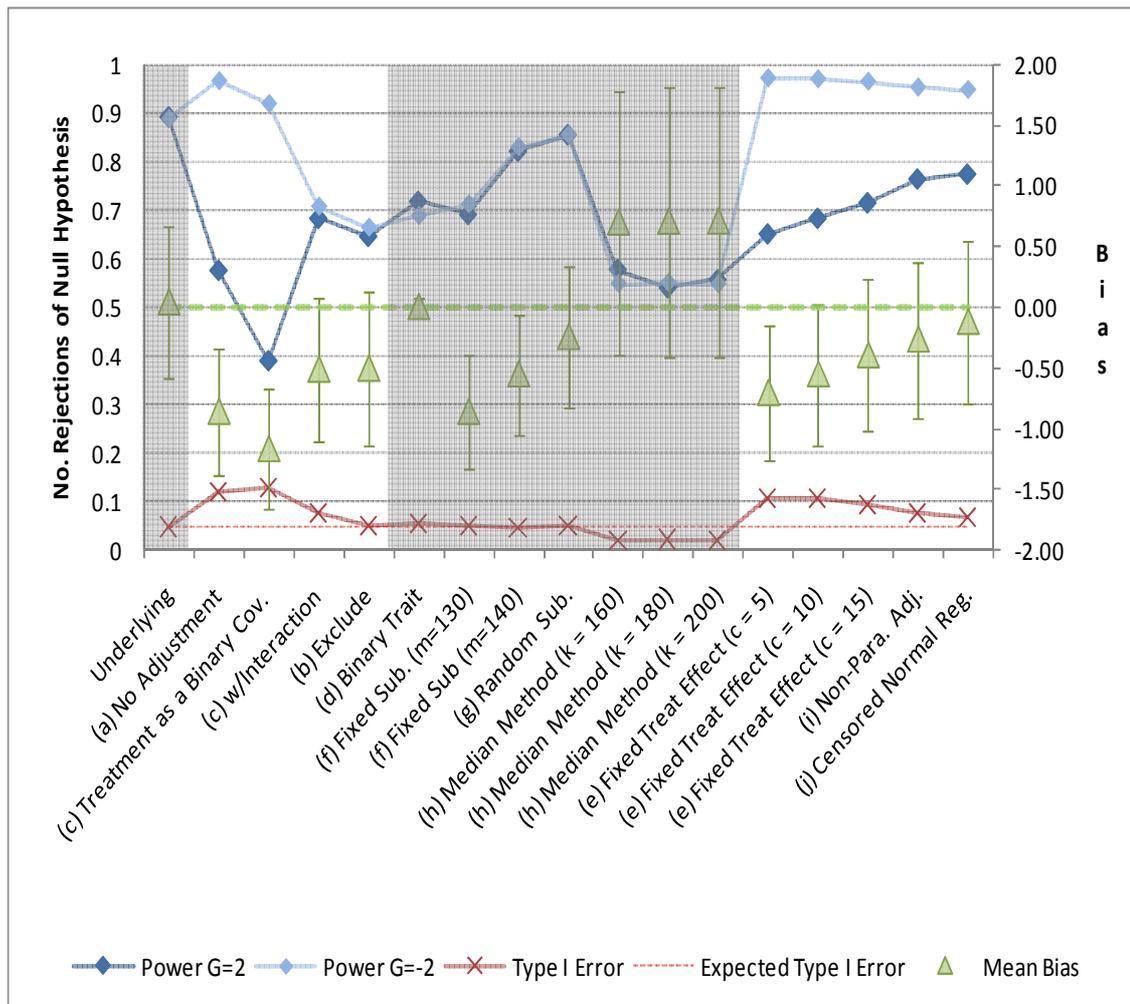


Figure 12: Graphical representation of the results for Scenario 6b. Approaches are arranged here in categories (from left to right): naïve, substitution, informative phenotype. Power to detect the genetic variant, g , when $\beta_3 = 2$ and $\beta_3 = -2$ are denoted in dark blue and light blue diamonds respectively, and type I error relative to the genetic effect is denoted in red crosses. Power and type I error are evaluated on the left vertical axis. Mean bias/SE with respect to $\beta_3 = 2$ is shown in green triangles, and is evaluated on the right vertical axis.

1.3.3 Scenario 7: Differential Probability of Receiving Treatment

In contrast to the previous two scenarios, which evaluate how the approaches perform when the treatment effect is differential, this scenario investigates the effect of a differential probability of receiving treatment. A differential probability (or *threshold*) of receiving treatment occurs where the chance of receiving

treatment depends on another factor other than the outcome of interest. As with the previous scenarios, this scenario also appears particularly relevant to studies of BP, because antihypertensives tend to be administered to individuals with diabetes at a lower threshold than to non-diabetics (Raum *et al.*, 2008). As such, a differential probability of receiving antihypertensive medication is known to occur in practice, and could affect real studies of BP.

Previous work has already assessed some of the approaches to analysis under a differential probability of treatment (as described in Section 1.1.5) (McClelland *et al.*, 2008). Although that paper finds that several of the approaches are impaired under this condition, the reported analyses focus on testing the approaches with respect to estimating the effect of the “differentiating factor” itself (in this case, diabetes). In the context of *this* work, a more realistic aim of the analysis is to identify possible genetic and/or lifestyle determinants of blood pressure (or other phenotypes), which will often be unlikely to directly affect the probability of receiving treatment themselves. The primary concern of these analyses, thus, is to detect the effects of “non-differentiating” factors. A more pertinent question of concern for the present work is therefore whether a differential probability of receiving treatment with one factor has knock-on effects on the estimates of any other factors. This scenario assesses the approaches under a condition where the threshold for receiving treatment depends on diabetes, but where the focus of the analyses is on estimating the effect of a genetic variant on BP.

1.3.3.1 Simulation Method

This scenario simulates a differential probability of receiving treatment by exposure to diabetes. Following the simulation procedure used for the General Simulation Study, underlying SBP is first generated from the linear regression model shown in Equation 12. Hence, each subject is generated an age (25-80), sex (male or female), and genotype for two independent diallelic loci with allele frequencies of 0.3 (for each genetic variant, a subject has 0, 1 or 2 copies of the minor allele). As before, one of the genetic variants assesses power and has an effect of +2 mmHg, while the other variant assesses type I error and has a null effect. For each individual, an indicator of diabetes is also generated. Diabetes is imposed randomly to subjects within each study with a prevalence of 20%. It has no effect on SBP.

This scenario allocates treatment in one of two possible ways. For hypertensives (i.e. those with an underlying SBP of at least 140 mmHg), treatment is allocated with a probability of 0.75 in the usual way. In addition, if an individual has diabetes and an SBP of at least 130 mmHg, treatment is allocated with a probability of 0.9. For those individuals allocated to receive treatment – by either means – a treatment effect is generated in the usual way, i.e. from the distribution used in the General Simulation Study. Hence, the treatment effect, $\gamma_i \sim N(15, 4^2)$.

Table 16 presents the full simulation properties for Scenario 7.

Parameter	Scenario 7
Simulation Runs	1000
Sample Size: n	2000
Intercept: β_0	110
Age (Years): age_i ; β_1	Uniform [25-80]; 0.4
Sex (M/F): sex_i ; β_2	Bernoulli (0.5); 3
Genetic Variant: g_{1i} ; β_3	Bin(2, 0.3); +2
Genetic Variant 2: g_{2i} ; β_4	Bin (2, 0.3); 0
Diabetes (Y/N): $diabetes_i$; β_5	Bernoulli (0.2); 0
Random Error	$\sim N(0, 18)$
Hypertension Criterion	$SBP \geq 140$
$P\{\text{Treated} \mid \text{Hypertensive}\}$	0.75
Hypertension with Diabetes Criterion	$SBP \geq 130$
$P\{\text{Treated} \mid \text{Hypertensive Diabetic}\}$	0.9
Treatment Effect (mmHg)	$\sim N(-15, 4^2)$

Table 16: Simulation Properties for Scenario 7.

Table 17 summarises the descriptive statistics for Scenario 7. As can be seen, a larger proportion of the sample is treated in this scenario compared with in the General Simulation Study (mean percentage treated = 33% here compared with around 28% in other scenarios). Just over half of the individuals with diabetes receive treatment in these studies, but only around a third of the total number of individuals on treatment are diabetic.

Summary Statistic		Scenario 7
Mean Underlying SBP (SD)	Overall:	133.70 (19.2)
	[Treat _i = 1]:	151.23 (11.4)
	[Treat _i = 0]:	125.00 (16.1)
Mean Observed SBP (SD)	Overall:	127.72 (15.8)
	[Treat _i = 1]:	136.23 (12.1)
	[Treat _i = 0]:	125.00 (16.1)
% SBP>140		37.14
% SBP>150		19.82
% SBP>160		8.53
% Individuals on Treatment		33.18
% Diabetics who Receive Treatment		54.52
% Non-diabetics who Receive Treatment		27.81
% Treated Individuals who are Diabetic		32.59
Mean Treatment Effect (SD)		15.00 (4.0)

Table 17: Descriptive Statistics for Scenario 7.

1.3.3.2 Results

Figure 13 below presents the results for Scenario 7 graphically, while the full table of results are shown in Table 53 of Appendix A. As can be seen, most approaches provide similar results in this scenario to those obtained in the General Simulation Study. The Informative BP approaches [*Fixed Treatment Effect* (e), *Non-Parametric Adjustment* (i), and *Censored Normal Regression* (j)] therefore perform well, with high powers (≈ 0.9) and generally accurate estimates of the parameter coefficients [e.g. mean $\hat{\beta}_3 = 2.08 - 2.23$ for (i) and (j); mean $\hat{\beta}_3 = 1.68 - 2.0$ for (e), when $c = 5, 10$ and 15].

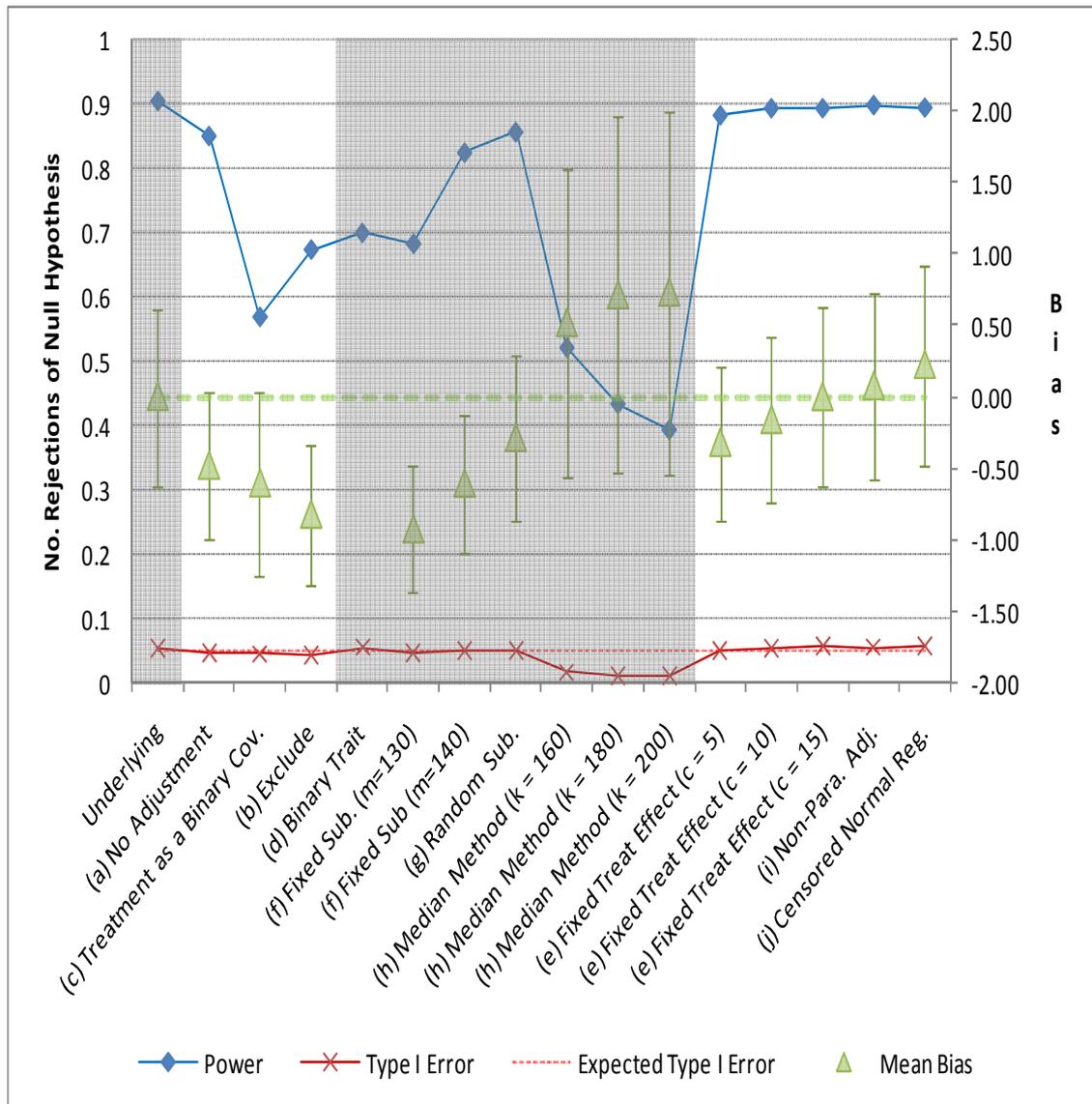


Figure 13: Graphical representation of the results for Scenario 7. Approaches are arranged here in categories (from left to right): Naïve, Substitution, Informative Phenotype. Power to detect the genetic variant g_1 (i.e. $\beta_3=2$) is denoted in blue diamonds, and type I error relative to the genetic variant g_2 (i.e. $\beta_4=0$) is denoted in red crosses. Both power and type I error are evaluated on the left vertical axis. Mean bias/SE with respect to $\beta_3=2$ is denoted in green triangles, and is evaluated on the right vertical axis.

Somewhat surprisingly, the Substitution approaches [*Binary Trait* (d), *Fixed Substitution* (f), *Random Substitution* (g), and the *Median Method* (h)] are relatively unperturbed in this scenario. These approaches assume that any

individual on treatment is hypertensive and, hence, could be expected to perform poorly here due to the allocation of treatment to some diabetics who are “normotensive”. With the exception of (h), which yields a small reduction of power here compared to other scenarios ($\approx 0.4-0.5$ here Vs around 0.5 elsewhere), none of the other Substitution approaches seems adversely affected here at all. A possible reason for this observation is that these approaches actually misclassify relatively few individuals in this scenario – despite the additional probability of receiving treatment for individuals with diabetes. For example, the only “misclassified” individuals here are diabetics with SBP between 130 mmHg and 140 mmHg. This represents, on average, only around 3.5% of the total sample. These “misclassified” individuals, furthermore, have relatively high BP, and the degree of misclassification imposed by the Substitution approaches can therefore be considered mild. Nevertheless, despite the fact that these approaches avoid suffering additional bias here, they remain sub-optimal approaches to analysis.

In support of previous findings (McClelland *et al.*, 2008), the full table of results (Table 53, Appendix A) shows that most of the approaches yield massively impaired estimates of the diabetes effect here. For example, diabetes has a null effect in this scenario, but estimates of its effect vary from approx. -7 mmHg [for (b) and (c)] to approx. 48 mmHg (!) [for (h)], often with high statistical significance (p-values not provided). The best approaches in terms of estimating the true effect of diabetes on SBP seem to be *Fixed Substitution* (f) and the Informative BP approaches [(e), (i) and (j)]. Nevertheless, in line with previous scenarios, (f) seems highly sensitive to its substitution parameter, m . Similarly, although (i) and (j) look acceptable here in comparison to other

approaches, they yield relatively high magnitudes of bias around the diabetes effect [e.g. mean bias \approx 1.65-1.85]. The only approach that truly performs well in terms of estimating the effect of diabetes is *Fixed Treatment Effect* (e). Assuming a fixed and appropriate size for the imputed treatment effect, this approach suitably adjusts for the use of treatment regardless of whether or not an individual is clinically hypertensive or normotensive.

Despite the biased estimates of the diabetes effect generally observed in this scenario, as stated above, the focus of an analysis will typically be on “non-differentiating” factors. Therefore, as the *Informative BP* approaches perform well in terms of estimating the effects of unrelated, independent parameters, they again appear the most appropriate methods to use in practice.

1.4. Discussion

Rapid progress is being made identifying genetic variants associated with BP in large-scale genome-wide association studies (Levy *et al.*, 2009; Newton-Cheh *et al.*, 2009). However, as yet unidentified genetic determinants of BP are likely to have even more modest effect sizes than those already discovered. As such, approaches to maximising the statistical power remain important, and the need for an appropriate approach to analysis – which controls type I error – remains vital.

1.4.1 Summary and Explanation of the Results

The simulations in Section 1.2 show that when the intervention is non-differential, the best approaches to analysis are clearly the *Informative BP* approaches [*Fixed Treatment Effect* (e), *Non-Parametric Adjustment* (i) and

Censored Normal Regression (j)]. The Informative BP approaches generally yield similar results to the optimal analysis of underlying BP in these settings, and, thus, appear to adequately control for the use of treatment. This finding concurs with previous work (Tobin *et al.*, 2005), which advises use of these approaches for analyses of BP.

The Informative BP approaches seem to perform well here because they exploit all the observed data within each analysis, and they therefore maintain the natural variability between BP measurements between individuals. The adjustments for treatment that they impose also seem appropriate for most realistic situations. For instance, these approaches seem relatively robust to situations where the distribution of the treatment effect is fixed (e.g. in the General Simulation Study and scenarios 1 and 2) and where the treatment effects depend on BP (e.g. when some individuals use combination therapy or when the treatment effect is proportional to the underlying BP – such as in scenarios 3 and 4 respectively). Conversely, the reason why the Informative BP approaches are affected by the differential treatment effects simulated in sections 1.3.1 and 1.3.2 is simply because they do not (and generally cannot) account for the SNP-treatment interaction. As these approaches utilise all observed data, they provide biased estimates of the marginal effect of any genetic variant that interacts with treatment because of the differential reduction in BP – by genotype – due to treatment. Although, in principle, *Fixed Treatment Effect* (e) could impose a simple adjustment for a pharmacogenetic interaction by imputing different treatment effects to different individuals (e.g. based on their genotypes), this would require prior knowledge of the interaction. To date, although there is strong evidence of a genetic component to the variability of BP

responses to antihypertensives, findings identifying loci for specific pharmacogenetic interactions have not replicated (Turner *et al.*, 2001; Arnett *et al.*, 2005; Arnett *et al.*, 2009). Hence, no firm data on the existence of pharmacogenetic interactions with antihypertensive medications currently exist, and any attempts to adjust for a pharmacogenetic interaction with (e) would be speculative.

In contrast to the Informative BP approaches, the Substitution approaches [i.e. *Binary Trait* (d), *Fixed Substitution* (f), *Random Substitution* (g) and *Median Method* (h)] and *Exclude* (b) are unaffected by the pharmacogenetic interactions implemented in scenarios 5 and 6. However, these approaches consistently yield sub-optimal results and, where applicable, seem highly sensitive to the values of the “substitution parameters”. Of the Substitution approaches, the *Median Method* (h) is particularly intriguing. Approach (h) assumes that anyone receiving treatment has an underlying BP greater than the median and, so long as fewer than half the participants receive treatment, *should* yield unbiased estimates of the parameter coefficients when this condition holds (White *et al.*, 2003). Nevertheless, neither of these conditions is ever violated in the majority of the scenarios tested in sections 1.2 and 1.3, yet (h) consistently overestimates the effects of the regression coefficients.

Section 1.1.2 explains that each of the Naïve approaches [*No Adjustment* (a), *Exclude* (b), and *Treatment as a Binary Covariate* (c)] is fundamentally flawed, and the results throughout sections 1.2 and 1.3 demonstrate this. The Naïve approaches consistently yield shrinkage bias in their estimates of all the regression coefficients, and, consequently, provide reduced statistical powers

compared to the Informative BP approaches. Although (a) sometimes yields a relatively high power, this approach is not immune to any of the conditions that cause problems for the Informative BP approaches and, hence, should never be used in favour of these approaches.

Related to the Naïve approaches, the additional analysis performed in scenarios 5 and 6 – which models the SNP-treatment interaction term – avoids the drastic biases due to pharmacogenetic interactions. This is because it accounts for the differences in treatment efficacy between individuals by fitting the interaction term. As stated earlier, this is the only approach that can easily account for a SNP-treatment interaction because it is the only model that includes treatment as a covariate. Nevertheless, despite avoiding these biases due to the pharmacogenetic interactions, this approach suffers from shrinkage bias and it yields a suboptimal power. It is now well established that modelling treatment as a binary covariate is a flawed approach to analyses of BP (Tobin *et al.*, 2005). For instance, because the use of antihypertensives in this setting both predicts BP and is a consequence of having high BP, treatment should not be handled as a conventional covariate. Doing so explains away variation within the data, and attributes this variation to an apparent “treatment effect”. Including a treatment main effect term within an analysis model can thus mask true causal factors of BP – such as genetic variants – which are usually the main focus of a study.

Given the above findings, there is no obvious choice of approach that can be expected to perform well in every situation. However, some practical recommendations are now discussed.

1.4.2 Practical Recommendations

It could be argued that because the Substitution approaches successfully control the type I error rates in all the scenarios performed in sections 1.2 and 1.3, their use should generally be preferred to the Informative BP approaches. However, as has already been noted, these approaches are highly reliant on the choices of the substitution values. Inappropriate choices of the substitution parameters could severely hamper an investigation in terms of its ability to detect any undiscovered genetic variants, which are suspected to have very small effect sizes. Although guidance on the choice of the substitution parameters is available (e.g. (White *et al.*, 1994; Hunt *et al.*, 2002; White *et al.*, 2003)), different values for these will be better suited to different circumstances. In practice, it may be difficult to choose suitable values for the substitution parameters – and, indeed, it would be difficult to verify how suitable existing choices are. A further limitation of the Substitution approaches is that they rely on the assumption that all individuals who use antihypertensives are hypertensive. This is clearly a strong assumption, because antihypertensive medications are sometimes prescribed for conditions such as coronary heart disease (CHD), heart failure and migraine. As the reason for their prescription may not be documented in a study, scenarios 2 and 7 specifically test for sensitivity to this assumption. None of the approaches are badly impaired in Scenario 7 because any *normotensive* subjects administered treatment here have relatively high BP (i.e. SBP between 130 and 140 mmHg). In contrast, Scenario 2 shows that the Substitution approaches perform poorly when the assumption is more seriously compromised, and this finding agrees with

previous work (Tobin *et al.*, 2005). For these reasons, I do not generally recommend use of the Substitution approaches for a primary analysis.

In addition to the results explicitly shown in this chapter, an additional scenario was also performed to investigate the effects of a pharmacogenetic interaction when the focus of the analysis is on estimating the effect of a genetic variant that is independent of the interaction. As has been shown in Section 1.3.1.2, estimates of the marginal effects of any independent genetic or non-genetic factors are unaffected if a pharmacogenetic interaction involves a different genetic variant. Hence, although the Informative BP approaches are affected in the presence of a pharmacogenetic interaction with the genetic variant of interest, they appear to remain the best approaches to use to estimate the main effects of variables that are uninvolved in such interactions. Scenario 7 also supports this observation. Here, estimates of the effect of diabetes – which is assumed not to be of interest – are impaired, but the Informative BP approaches remain unaffected in terms of their estimates of the effects of all other regression terms. Note, however, that these conclusions apply to estimates of the effects of independent variables only, and estimates of the effects of genetic variants that, for example, are correlated (i.e. in *linkage disequilibrium*) with a SNP involved in a pharmacogenetic interaction will also be affected using the Informative BP approaches. Similarly, in Scenario 7, the Informative BP approaches could be expected to provide biased estimates of the effects of any variants that are associated with diabetes. Further work is clearly required to investigate the possible extent of these biases on estimates of the effects of variables correlated with a “differentiating factor” in other realistic settings.

In an ideal world it would be possible to identify a priori those SNPs likely to be involved in differential treatment effects (for example, from published pharmacogenetic studies of BP). Tests of the marginal effects of these particular SNPs could then be performed with an approach immune to the effects of a pharmacogenetic interaction (such as one of the Substitution approaches), while tests of all remaining, independent SNPs could use the Informative BP approaches. As noted above, however, no firm knowledge about pharmacogenetic interactions with antihypertensives currently exists. Nonetheless, it seems reasonable to assume that only a small proportion of genetic variants across the human genome will alter the efficacy of antihypertensive treatments. Given this assumption, a reasonable recommendation to make is that primary analyses of BP – which aim to detect genetic variants that have an effect on underlying BP – should be performed using the Informative BP approaches. Any results require a critical interpretation, however, due to the lack of information about which regions of the genome have discernable effects on underlying BP and also alter the efficacy of antihypertensives.

Pending further firm biological evidence about pharmacogenetic interactions, there may be exploratory analyses that could be undertaken with a dataset under study to provide insight about potential interactions with a genetic variant of interest. Although not recommended as a primary analysis, one approach to investigating the possible presence of a pharmacogenetic interaction for a variant of interest would be to use the extended analysis of *Treatment as a Binary Covariate* (c) [i.e. which models the SNP-treatment interaction term] to explicitly test for the interaction. Interactions are generally detected at a lower

power than main-effects, however, and extensive follow-up work is required to clarify whether such an approach would be reliable. A possible alternative approach to identifying a pharmacogenetic interaction could be to compare findings from an Informative BP approach and one of the Substitution [or *Exclude* (b)] approaches. The latter approaches are unaffected by pharmacogenetic interactions. If the results from the two analyses do not differ substantially, it may be reasonable to assume that no strong pharmacogenetic interaction is present. However, further work is again required to illustrate how large a discrepancy between the findings of these different approaches might be expected for real situations, as evidence about the characteristics of variants (e.g. minor allele frequency, main effect and interaction effect sizes and directions) involved in pharmacogenetic interactions becomes available.

1.4.3 Implications

Section 1.3 shows that otherwise sensible approaches to the analysis of BP are affected when a genetic variant of interest influences treatment efficacy. Estimates of the marginal effects of genetic variants involved in pharmacogenetic interactions may therefore be biased – possibly leading to false-negative and false-positive findings. Pharmacogenetic interactions can thus impact on the statistical power of a study and on the level of type I error.

In principle, these results suggest that the reported findings from existing genetic association studies could contain errors as a result of pharmacogenetic interactions. For instance, a genetic variant that influences treatment efficacy could yield spurious association with BP, or, conversely, a genetic variant that truly influences BP could be masked if it is also involved in a pharmacogenetic

interaction. A secondary aim of this work could be to characterise such cases. Although analyses such as the *Binary Trait* approach (d) are low powered for a primary analysis, they could provide useful subsequent checks to help identify whether novel genetic associations could be driven by a pharmacogenetic effect. For instance, all the genetic variants reported by Newton Cheh *et al.* (Newton-Cheh *et al.*, 2009) were associated with dichotomous hypertension in addition to continuous SBP and DBP, and are therefore unlikely to be fallacious. The issue of type I error due to a pharmacogenetic interaction is thus unlikely to be a problem in this particular study. However, the possibility of *type II error* remains. In addition to the strength of the interaction and the number of individuals involved, type II error will also depend on the direction of the interaction in relation to the direction of the main effect.

1.4.4 Applicability of the Findings

This chapter solely uses simulation to demonstrate the potential influences of different realistic conditions because, in practice, the true model generating mechanism is unknown. For instance, as yet, there is little known regarding the true nature and magnitude of pharmacogenetic interactions with antihypertensives. The true influence of pharmacogenetic interactions in real analyses of BP is therefore difficult to determine. If particular genetic variants interact with multiple classes of antihypertensive, there is a potential for serious distortions of the data (such as those shown in Scenario 5), but if pharmacogenetic interactions are specific to particular classes of antihypertensive, the implications could be less drastic (such as in Scenario 6).

Until now, this chapter has focussed solely on analyses of BP, but the findings are also relevant to the analysis of other traits. For example, cholesterol-lowering drugs are widely used within western countries, and the investigation of low-density lipoprotein (LDL) and high-density lipoprotein (HDL), thus, may also require one of the corrections for treatment described. Notably, because a single class of treatment – statin therapy – is predominantly used to lower cholesterol, the findings from Scenario 5 may be especially relevant to these traits. For instance, any pharmacogenetic interaction would most likely apply to the majority of subjects on treatment and, hence, the conditions simulated in Scenario 5 (which may be considered extreme for a study of BP) may be quite typical of a study of LDL/HDL.

Ultimately, both forms of a differential intervention simulated in Section 1.3 lead to similar conclusions. For instance, estimation of the parameter that modifies either the treatment effect or the threshold for receiving treatment is often distorted, but estimation of all other (independent) parameters is generally unaffected. Hence, if the “differentiating parameter” itself is known but is not of interest, analyses may be performed without regard to these findings; however, when the modifying parameter needs to be estimated (and may or may not be unknown), difficulties may arise. Although Section 1.4.2 suggests possible approaches to verifying results from genetic analyses of BP and to identifying potential pharmacogenetic interactions, further work is clearly required in these areas.

1.4.5 Conclusions

Consistent with previous findings (Tobin *et al.*, 2005), the work in this chapter suggests that the Informative BP approaches remain the most reasonable approaches to use for primary analyses of the main effects of SNPs in most settings. Nevertheless, Section 1.3 demonstrates that caution is required in the interpretation of any associations obtained from these approaches. If there is strong *a priori* evidence of a particular pharmacogenetic interaction – or of genetic variants associated with a factor that influences the probability of receiving treatment – it makes sense to consider the results of a different approach for the particular genetic variants involved. As further evidence of the nature and magnitude of pharmacogenetic interactions with BP emerges, more detailed examination of the various approaches, their comparability, and possible methods for checking for these interactions will be warranted.

Chapter 2.

Participant Identifiability in GWAS

2.1. Introduction

Data collected in genetic epidemiological studies are, by nature, extremely sensitive, and steps must be taken to ensure the protection of participant confidentiality. As such, there are strict laws to govern the sharing of individual level genetic and non-genetic information (Lowrance *et al.*, 2007). Advances in genomic research are, however, informed and accelerated by the publication of results and summary information from genetic epidemiological studies (McCarthy *et al.*, 2008), and, indeed, the sharing of these data has often been a condition demanded by funding bodies (Kaye *et al.*, 2009). But in 2008, a statistical method was published that potentially allows individuals to be identified in genome-wide association studies (GWAS) using only summary information (such as allele frequencies) often freely available on the Web (Homer *et al.*, 2008). In response to the publication of this method, the National Institutes of Health (NIH) and the Wellcome Trust were compelled to alter their guidelines on the release and publication of summary information from GWAS and, consequently, they withdrew summary data such as allele frequencies from the internet (see Appendix B.1) (Couzin, 2008). Access to summary data from GWAS has since become restricted only to registered and approved

researchers. Hence, rather than moving towards the scientific ideal of open, public access to genomic data (Smith, 2009), the culture of data sharing in genomics research is currently heading the opposite way.

Since the publication of the above method, uncertainties surrounding its true implications have been rife. For instance, there has been some debate as to whether or not the method (which shall be referred to here as the “Homer” method) can be used to identify any person, in any study, under any circumstances. In order to cast light on such speculation, a full understanding of the science behind the method is required. This work therefore aims to examine the Homer method in greater depth.

There are two main perspectives from which to approach these investigations. The Homer method was originally proposed from a forensic perspective as a means by which to identify individuals from pooled DNA samples such as found at crime-scenes. The forensic perspective therefore seeks to understand the Homer method in order to clarify how it can be used, in what circumstances it can be used, and what its limitations are. In contrast, the genomics community aims to avoid the risk of participant identification. Genomics researchers therefore want to know how to prevent participants from being identified, and what to do in spite of this threat to participant confidentiality posed by the Homer method. This work aims to shed light on these issues by considering the problem from both perspectives.

Section 2.2 introduces the Homer method and discusses some important practical considerations for its use. A review of the relevant literature that has

been published since the Homer *et al.* (Homer *et al.*, 2008) paper is provided in Section 2.3, detailing the implications of the method and some extensions to the approach. I empirically test the Homer method – via simulation – in Section 2.4, to demonstrate how it potentially performs in practice.

Subsequent sections then focus on a particular alternative approach to the Homer method, which can also potentially test for presence in GWAS using SNP allele frequencies. Section 2.5 describes this alternative approach, before it is tested via simulation in Section 2.6. Implications of the method's use in real data are discussed in Section 2.7, and some further modifications to the test are proposed. Section 2.8 investigates sensitivity of the method to its core assumption of co-ancestry (see Section 2.2.4). Finally, Section 2.9 considers what can be published from GWAS given these findings, before a general discussion is provided in Section 2.10.

2.2. The Homer Method

I first describe the Homer method in accordance with the original paper (Homer *et al.*, 2008) in Section 2.2.1, before subsequent sections provide a critique. In particular, Section 2.2.2 focuses on practical issues with its use, Section 2.2.3 explores its different possible applications, Section 2.2.4 describes its assumptions, and Section 2.2.5 discusses any other important considerations for its use.

2.2.1 Outline of the Method

The Homer method (Homer *et al.*, 2008) has been proposed as a means by which to identify whether a particular individual contributed to a mixture of DNA

(which is a genetic sample containing DNA from multiple individuals). The method was originally proposed for use in a forensic setting, whereby the mixture would be a DNA sample obtained from a crime-scene and the individual of interest would be a suspect for the crime. Summary statistics published from genetic epidemiological studies can also be considered to denote the characteristics of a mixture of DNA, however (see Section 2.2.3). This section outlines the original method, before subsequent sections consider its application in the GWAS setting, in addition to the forensic context.

For a particular single nucleotide polymorphism (SNP), the Homer method compares an individual's scaled genotype (see next paragraph) to the allele frequency in the mixture and to the allele-frequency within some known reference population (which is to be described in Section 2.2.2). Over many SNPs, it forms a "distance" metric, which determines probabilistically whether an individual is "closer" to the mixture or "closer" to the reference population. If the individual of interest is statistically significantly closer to the mixture, then, in theory, presence of the individual within the mixture can be inferred due to the large number of SNPs typically involved. The Homer method can be expressed as follows.

For the i^{th} individual of interest, Y_{ij} denotes the scaled genotype for the j^{th} SNP ($j = 1, \dots, s$). Typically, Y_{ij} has a scaled value of 0, 0.5 or 1 (representing 0, 1 or 2 copies of the minor allele respectively), but, alternatively, Y_{ij} can be a measure of probe intensity taking a continuous value between 0 and 1. Note that Y_{ij} is thus an observed allele frequency for individual i . For a known reference population (to be discussed in Section 2.2.2), the minor allele frequency (MAF)

for the j^{th} SNP is denoted Pop_j , and for the mixture, the MAF for the j^{th} SNP is denoted M_j . Assuming that the reference population and the mixture are ancestrally similar (i.e. they have similar allele frequencies across all SNPs), the distance measure, $D(Y_{i,j})$, is defined as:

Equation 14

$$D(Y_{i,j}) = |Y_{ij} - Pop_j| - |Y_{ij} - M_j|$$

Under the alternative hypothesis that the individual of interest is in the mixture (H_1), the individual's presence in the mixture drives the test and, hence, over s SNPs, mean $D(Y_{i,j})$ will be greater than zero. This is based on the principle that if an individual is part of a sample, his/her genotypes will be "closer" to the sample means – that is, to the allele frequencies in the *mixture* – than to the population means. Under the null hypothesis that the individual is *not* in the mixture (H_0), Homer *et al.* state that "a random individual should be equally distant from the mixture and the mixture's reference population". Under H_0 , mean $D(Y_{i,j})$ is thus assumed to be zero. The validity of this assumption is examined in greater depth in the following section.

To test mean $D(Y_{i,j})$ for departure from zero and, hence, to test the null hypothesis in view of rejecting it in favour of the alternative hypothesis, the authors propose using a one-sample t-test. A one-sample t-test is defined as

$$\frac{\bar{x} - \mu_0}{d/\sqrt{n}} \sim t_{n-1},$$

where d is the sample standard deviation and n is the number of observations.

The test-statistic, $T(Y_i)$, is therefore

Equation 15

$$T(Y_i) = \frac{\frac{\sum_{j=1}^s D(Y_{ij})}{s} - \mu_0}{\sqrt{\text{Var}[D(Y_{ij})]/s}} \sim t_{s-1},$$

where i , j and s are as before, and where μ_0 is assumed to be 0 (i.e. because under the null hypothesis a random individual is assumed to be equidistant to the mixture and the reference population). The variance, $\text{Var}[D(Y_{ij})]$, is simply the sample variance:

$$\text{Var}[D(Y_{ij})] = \frac{1}{s-1} \sum_{j=1}^s \left(D(Y_{ij}) - \frac{\sum_{j=1}^s D(Y_{ij})}{s} \right)^2.$$

In the following sections, I examine some of the practical issues and some of the problems in using the Homer method as stipulated above.

2.2.2 Practical Issues

Before the Homer method can be applied in practice, there are a number of issues that first need to be considered. This section introduces these issues and discusses their possible implications.

2.2.2.1 Reference Population

As we have seen, the Homer method was originally proposed as a new forensic test to infer presence within a mixture of DNA using single nucleotide polymorphisms (SNPs). In contrast, most existing forensic tests are based on a small number of short tandem repeats (STRs). Because existing tests only ever use 13-15 well studied STR markers, population frequencies for these STRs are known with a good degree of certainty. However, the Homer method suggests

that vast numbers of SNPs may be required; for instance, up to 500,000 SNPs or even more. Due to the large number of SNPs required, and the fact that SNPs are not as well studied as forensic STRs, precise estimates of population allele frequencies are difficult to obtain. Thus, population allele frequencies for SNPs (which are denoted Pop_j in Section 2.2.1 above) are not currently available, and, although they could become available in future, use of the Homer method in practice will require estimating them. Realistically, this can only be achieved by sampling a “reference group” from the reference population, and using the allele frequencies in the reference group as estimates of Pop_j . In practice, the Homer method is therefore really a two sample problem (i.e. which compares an individual of interest to a reference group and to a mixture) (Jacobs *et al.*, 2009), rather than the one sample situation originally described (i.e. where the mixture is the only sample because true allele frequencies in the reference population are assumed). Throughout this chapter I thus use the two terms “reference group” and “reference population”, and it is important to note that these are distinct.

The original Homer *et al.* paper does not go into details on the reference population, but this two sample situation actually seems better suited to the proposed test statistic than the one sample situation originally proposed. For instance, in the one sample situation, if the test individual is not in the mixture, he/she must be part of the reference population (assuming that the assumption of co-ancestry between the reference population, mixture, and the individual of interest applies – see Section 2.2.4). Under the null hypothesis, the individual, thus, will be likely to be “closer” to the reference population than to the mixture and, hence, the test statistic, $T(Y_i)$, will be *less* than zero under the null

hypothesis, rather than *equal* to zero as assumed. In the two sample situation the assumption that $T(Y_i)$ will be zero under the null hypothesis seems more reasonable. For instance, if an individual is not in the mixture, then as long as he/she is not in the reference group either, it seems logical that he/she could be equidistant to both groups. From here on, I will therefore focus on the two sample problem, but this issue will be discussed again in Section 2.4. ,

2.2.2.2 Composite Hypotheses

The use of the Homer method in a two sample setting has important implications for the hypothesis testing procedure. A one sample test (i.e. which uses population allele frequencies) allows, in principle, an alternative hypothesis that an individual of interest is in the mixture to be tested against a null hypothesis that he/she is *not* in the mixture (and is, hence, a random member of the population). In contrast, the two sample problem tests a different set of hypotheses. Because two groups are compared, the Homer method can in theory determine whether an individual is in one group or the other (or, additionally, whether he/she is in neither group). For instance, if $T(Y_i)$ is significantly greater than zero the test would imply that the individual is in the mixture, and if $T(Y_i)$ is significantly *less* than zero the test would imply that the individual is in the other group. In the following section I describe some implications of this two-sample test in the context of a case-control GWAS. The original purpose of the test, however, is in a forensic context, and from this perspective the interest is only ever to determine whether an individual is in the mixture or not. Under the alternative hypothesis, $T(Y_i)$ must always be greater than zero to infer that the individual is in the mixture. However, under the null

hypothesis (i.e. that the individual is *not* in the mixture), $T(Y_i)$ could be less than zero (to infer that he/she is in the reference group) or equal to zero (to infer that he/she is in neither group). The null in this context is therefore a composite hypothesis, in which the distribution of the test statistic, $T(Y_i)$, is not properly specified. The validity of composite hypotheses has been questioned; Sir Ronald Fisher, for instance, has stated that a null hypothesis “must be exact, that is free of vagueness and ambiguity, because it must supply the basis of the ‘problem of distribution,’ of which the test of significance is the solution” (Fisher, 1966). By considering how frequentist “tests of significance” are derived, it is clear why composite hypotheses are problematic. For example, p-values – which are typically used to denote the statistical significance of a test – are defined as: “the probability of obtaining a result at least as extreme as the observed result under the null hypothesis”. Hence, if there is no exact definition of the distribution under the null hypothesis, p-values cannot be calculated in the usual way.

Despite the above problems, it could be argued that the Homer test actually benefits from having a composite null. For instance, the assumption that $T(Y_i) = 0$ under the null hypothesis represents a “worst-case” scenario because, as explained above, $T(Y_i)$ for an individual who is not in the mixture is actually expected to be less than zero. Hence, under the null, the actual (or *empirical*) value of $T(Y_i)$ at any given quantile is expected to be less than the theoretical value (i.e. which is based on the standard normal distribution and which denotes the statistical significance of a test). There will, thus, be less type I error than ordinarily expected. Conversely, however, another implication of this composite hypothesis is that, under the alternative hypothesis, the power of the

test will be sub-optimal. For instance, the distribution of $T(Y_i)$ is expected to be centred above zero under the alternative hypothesis, so the “distance” between this and the theoretical null distribution (centred at zero) will be less than the distance between this and the empirical null distribution (centred below zero). The power to discriminate between the alternative hypothesis and the theoretical null hypothesis is thus less than it would be to discriminate between the alternative and the empirical null distribution. The composite hypothesis in this situation therefore leads to a conservative test: yielding less type I error at the expense of a lesser power. The simulations in Section 2.4 further investigate the type I error rates and power of the Homer method, and demonstrate these relationships in graphical displays (e.g. see Figure 16 and Figure 17 on pages 137-139).

We have so far seen that the Homer method, *in practice*, is a two sample test rather than the one sample problem originally stipulated, and that this has implications to the hypothesis testing procedure. In the following section, I outline a number of possible applications of the test, and I discuss how the hypothesis testing procedure varies for each.

2.2.3 Different Applications

There are several potential applications of the Homer method. This section describes what these different applications are and what effects they have, if any, upon the method’s testing procedure.

(a) Forensic

The forensic application of the Homer method – whereby the mixture is a DNA sample obtained from a crime-scene and the individual of interest is a suspect for the crime – has already been discussed. However, the method's suitability for use in this situation is dependent on a number of factors. Because the population frequencies for forensic STR markers are available, existing forensic tests are able to test the alternative hypothesis that the suspect is in the mixture against a null that specifically states that he/she is a random member of the population. As population allele frequencies are not available for SNPs at present, it is debatable whether the Homer method can truly examine the same null hypothesis as other STR-based forensic tests. For instance, although the Homer method may be able to conclude that an individual is not in the mixture, this is not necessarily the same as concluding that he/she is a random member of the population.

(b) GWAS Cohort

Testing for presence in a GWAS cohort would involve a similar use of the Homer method as an application in forensics. For example, where the forensic mixture is a DNA sample obtained from a crime, here, the mixture is a GWAS cohort and, instead, the DNA from an individual of interest may have been recovered from a crime. Crucially, an implication of using a GWAS cohort rather than a forensic mixture is that, in GWAS, all individuals contribute equally to the allele frequencies, whereas this is not guaranteed in a forensic mixture.

Use of the Homer method in this context would involve testing an individual against the allele frequencies from a GWAS cohort and against the allele frequencies in a reference group (compared to a genomic mixture and a

reference group in the forensic application). A hypothesis test for this setting would ideally test a null that an individual is “not in the cohort” versus an alternative that he/she is “in the cohort”. However, under this null hypothesis, the test statistic, $T(Y_i)$, has no specific distribution because an individual could be in the reference group or in neither of the two test groups if he/she is “not in the cohort”. Acknowledging this issue, an application of the method in this context could test the hypotheses:

H_0 : Neither group [$T(Y_i) = 0$]; Vs

H_1 : In the study [$T(Y_i) > 0$],

in a one-tailed test, assuming that a particular test individual would be unlikely to be in the reference group itself (and that this outcome would not be of interest anyway).

The potential threat that the Homer method poses toward the identification of participants from GWAS is what prompted the NIH and the Wellcome Trust to remove GWAS allele frequency data from the Web (e.g. see Appendix B1). For instance, any positive test result indicating that an individual participated in a particular study would breach most participant confidentiality agreements. Furthermore, if the DNA sample from the individual of interest had been recovered from a crime scene, a positive test result inferring that the individual participated in the study could lead to the authorities demanding the release of the full individual level data from the cohort. This application of the Homer method is one of the main focuses of this work, and subsequent sections focus on this application of the method further.

(c) Case-Control GWAS

Another potential application of the Homer method is to utilise the published allele frequencies from a case-control GWAS in an attempt to identify an individual of interest. In this context, a reference group would not be required because the case and control groups would instead be tested directly against one another. The test in this setting would have three possible outcomes, because an individual of interest must be a case, a control, or neither. In theory, however, statistical hypothesis testing can only ever discriminate between two hypotheses at a time. It thus seems necessary to have to perform the test twice to discriminate between the three hypotheses (comparing one pair of hypotheses each time). The test hypotheses must therefore be reformulated accordingly. In principle, there are different ways in which to formulate the test hypotheses but, as we shall see, each formulation has particular limitations.

The obvious way to formulate the test hypotheses is to simply test for presence in the case group first (using the controls as a reference), before testing for presence in the control group (using the cases as a reference). However, formulating the hypotheses as:

Test 1a: H_0 not a case Vs H_1 case; and

Test 2a: H_0 not a control Vs H_1 control,

involves testing composite null hypotheses. For example, “not a case” implies that mean $T(Y_i) \leq 0$. Hence, an alternative formulation of the test hypotheses that avoids the problem of a composite null could be:

Test 1b: H_0 not in study Vs H_1 case; and

Test 2b: H_0 not in study Vs H_1 control.

Potentially major problems with the above hypothesis formulation remain, however. Classic frequentist statistics stipulate that the null hypothesis can never be *accepted*, because data or evidence are only ever deemed to be consistent (or inconsistent) with the null distribution. Yet, in the above formulation of the test hypotheses, the conclusion that the individual of interest is not in the study cannot be reached without accepting the null hypothesis for both tests. A further issue with the above hypothesis is that neither Test 1b nor Test 2b accounts for all eventualities. For instance, the test of “not in study vs case” ignores the possibility that an individual could be a control, and the “not in study vs control” test ignores the possibility that an individual could be a case. Hence, one of the above tests will lead to an incorrect conclusion if the individual of interest is in the study regardless of whether or not the null hypothesis is rejected.

To counter this problem, a better formulation of the test for case-control GWAS data involves a *two-tailed* hypothesis:

Stage 1: H_0 : not in study [hence $T(Y_i) = 0$] Vs H_1 : in study [$T(Y_i) \neq 0$].

If the above null hypothesis is rejected, the evidence would suggest that the individual is present in the study. Subsequently, a second test would have to be performed to ascertain which of the two groups the individual is in:

Stage 2: H_0 : control [$T(Y_i) < 0$] Vs H_1 : case [$T(Y_i) > 0$].

The difficulty in adopting this strategy concerns how to implement the Stage 2 test, as each of the hypotheses in this test is composite (i.e. each involves a non-exact distribution). Although it may be possible to estimate the expected distribution of $T(Y_i)$ under each of these hypotheses (e.g. by simulation or by using an additional dataset) in view of performing a hypothesis test based on these specific distributions, this approach could also be problematic. For example, for a given number of SNPs and for the particular sample sizes of the two test groups (which are both known in advance), a putative null distribution for an individual being a control might be estimated to have a non-central t-distribution, for example, with a mean of -10 and a variance of 1. Thus, hypothetically, if the observed test statistic is -5, the test would lead to the individual being inferred to be a case. However, rejecting the null hypothesis here (and, hence, inferring the individual to be a case) is clearly nonsensical, because negative values of the test statistic actually imply that an individual is “closer” to the control group than to the cases. This issue highlights a limitation of traditional statistical methods for the testing of composite hypotheses. Classical (or *frequentist*) hypothesis testing procedures may be inappropriate for this class of test, and a completely satisfactory formulation of the test hypotheses for the Homer method in case-control GWAS data is thus difficult to achieve.

Out of each of the formulations of the test hypotheses stated above, the only completely satisfactory test is the two-tailed test of the null hypothesis that the individual is not in the study (*Stage 1* above). In practice, this therefore seems the most appropriate formulation of the test to use. However, as already discussed, a subsequent test (i.e. to assess whether the individual is a case or

a control) cannot easily be performed formally. Nevertheless, as the distribution of $T(Y_i)$ is expected to be distinct for individuals in each study group, a possible way around this problem may be to avoid formally testing for case-control status altogether. Instead, it may be possible to discriminate between case and control status based only on an informal check. For instance, any individual in a case-control GWAS **must** either be a case or a control and, hence, if the two-tailed test yields a significant result, the *sign* of $T(Y_i)$ alone should be a reliable indicator of case-control status. This strategy is investigated further in later sections (e.g. sections 2.4 and 6).

A final formulation of the Homer test for case-control GWAS data could therefore test the following hypotheses:

H_0 : Not in study [$T(Y_i) = 0$]; Vs

H_1 : In study [$T(Y_i) \neq 0$]

where, if H_0 is rejected, accept H_{1a} : *Case* if $T(Y_i) > 0$ or accept H_{1b} : *Control* if $T(Y_i) < 0$, acknowledging that discrimination between H_{1a} and H_{1b} does not require a formal test of significance when H_0 is rejected.

(d) Case or Control?

A final application of the Homer method is to infer case or control status for an individual already known to have participated in a case-control GWAS. This situation, thus, is similar to (c), but does not require a *Stage 1* test, i.e. to test for presence in the study, because the individual of interest is already a known participant. In this scenario, the null hypothesis would therefore assume that an individual is in one of the case or control groups, while, under the alternative hypothesis, the individual would be in the other group.

This scenario would be unlikely to occur in practice but, nevertheless, its potential to breach participant confidentiality agreements in GWAS needs to be considered. For example, if it were somehow known that an individual participated in a study, one could attempt to use the Homer method maliciously to ascertain whether the individual is a control or a case (and, hence, whether he/she has a particular disease associated with that group).

As discussed in (c), a simple formulation of the hypotheses:

H_0 : Control Vs H_1 : Case,

is not straightforward to test because both hypotheses are composite. It is thus difficult to formally define this test. In practice, however, this test could simply be performed in the same way as the two-tailed test described in (c). The one difference between the test here and in (c) would be in how to interpret non-significant results: here, a failure to reject the null hypothesis would imply that

there is insufficient evidence to infer case-control status, while in (c), failure to reject H_0 implies that the individual is not in the study.

Although, as already stated, an application of the Homer method in this scenario is not particularly likely in practice, a possible extension of its use to a similar scenario could have potentially major future ramifications. For instance, if a similar distance metric could be developed to infer disease status without necessarily requiring an individual to be present in the actual case or control groups used in the test, this would have major implications. Similarly, a future test based on similar principles to the Homer method may be able to infer a test individual's ethnicity if the two groups used in the test are of different ethnicities. In its present form, the Homer method is unlikely to have sufficient power to reach such conclusions for individuals in neither of the test groups; however, in future such tests may become tractable (e.g. with the sequencing and release of even larger proportions of the genome).

2.2.4 Assumptions

The Homer method assumes both explicitly and implicitly a number of conditions that may have important practical implications for its performance. For instance, a key, explicit assumption of the Homer method is that the reference population and the mixture are of similar ancestry. Homer *et al.* (Homer *et al.*, 2008) also state that it is “obvious” that the reference population must also be either “accurately matched in terms of ancestral composition to... the person of interest”, or “limited to analysis of SNPs with minimal (or known) bias towards ancestry”. This shall be referred to throughout as the assumption of “co-ancestry” – meaning that all individuals are from the same population (or

gene-pool). Co-ancestry between the individual of interest, the mixture, and the reference population is a required condition because it ensures that a test is only influenced by the individual's presence in or absence from the mixture. In GWAS, the ancestry of a particular cohort is usually well known, so choosing a well-matched reference group for a particular mixture should be relatively straightforward. Similarly, the two groups of a case-control GWAS are usually well matched in terms of ancestry, so an application of the method in this respect may not be problematic. Nevertheless, sometimes an individual of interest may have unknown ancestry or, in a forensic application of the test, the ancestry of the mixture may be unknown. Therefore, in some situations it may be difficult to ensure that the co-ancestry assumption is upheld.

Because the Homer method is based on a one-sample t-test, independent observations are assumed and the distance measures, $D(Y_{i,j})$, must be independent across all s SNPs used in the test. However, in practice, alleles for SNPs located closely to one another across the genome are more likely to be inherited together (this association is known as *linkage disequilibrium*). Where many SNPs are used to perform the Homer method, linkage disequilibrium (LD) will be an issue because it will violate this assumption of independent observations. However, if an adequate power can be achieved using fewer SNPs, it may be possible to only use SNPs that are widely dispersed across the genome (and, hence, which are not in linkage disequilibrium). Correlated observations are problematic in many statistical tests because they are less informative than an equivalent number of independent observations. The variance of a statistic based on observations that are correlated will thus be greater than one based on an equivalent number of independent observations.

Where the Homer method assumes independent observations, if the SNPs used in the test are, in fact, in LD, the calculated variance of the distance metric will be biased downwards, and this could lead to false-positive findings.

2.2.5 Other Important Characteristics

A number of characteristics affect how the Homer method performs. For example, the power of the test is influenced both by the number of (independent) SNPs available to use and the proportion of DNA contributed to the mixture by the individual of interest. Homer *et al.* (Homer *et al.*, 2008) report a simulation study investigating the trade-off in power between these two factors. When the proportion of DNA contributed to the mixture by the individual of interest is 0.1, they report that approximately 1,000 SNPs are required to identify the individual at a p-value less than 10^{-6} . For proportions of 0.01 to 0.001, approximately 10,000 to 25,000 SNPs are required (also at p-val < 10^{-6}).

In addition to its influence on the statistical power of the test, recent work by Egeland *et al.* (Egeland *et al.*, 2010) suggests that the proportion of DNA contributed to the mixture by the individual of interest can also affect the error rate. For instance, incorrect conclusions are reported when the proportion of DNA contributed to the mixture by each individual differs. In GWAS, all participants contribute equally to the allele frequencies in a particular cohort, and the proportion of DNA contributed by the individual of interest to the mixture is simply equal to the inverse of the study size (which is known). This finding, thus, is not applicable to GWAS applications of the method. In forensics, however, the proportion of DNA contributed to the mixture by the individual of interest would be unknown. Hence, the Homer method may be of limited use in

forensic applications, because the proportion of DNA contributed to the mixture by each individual cannot be guaranteed to be equal. Although this finding seems to contradict the results reported by Homer *et al.* (i.e. shown above), the original paper is unclear about whether or not each individual contributes equally to the mixture of DNA. Given the findings of Egeland *et al.*, it is possible that Homer *et al.* only simulated mixtures in which all individuals contribute equally.

The Homer *et al.* simulations also consider the effect of genotyping error on the performance of the method. In large mixtures, high degrees of genotyping error result in a marginally reduced power to identify individuals, but genotyping error has negligible effect in smaller mixtures.

Minor allele frequency (MAF) is another factor that can influence how the Homer method performs. For instance, SNPs with an especially low MAF will usually contribute little information to the test and, as such, could result in the requirement of a greater number of SNPs. Furthermore, genotyping error tends to be more common for SNPs with rare alleles. Homer *et al.* therefore recommend substituting any SNPs with minor allele frequency less than 0.05 for other SNPs.

A further factor that can influence the performance of the Homer method is whether any relatives of the individual of interest are included in either the reference group or the mixture. Using real data, Homer *et al.* show that where a mixture contains a first-degree relative of the individual of interest, the individual can falsely be inferred as present in the mixture. In this situation, the power of the test is typically reduced by half – which represents the proportion of alleles

shared, on average, between any individual and a first-degree relative (i.e. a parent, child or sibling). In practice, it may thus be difficult to confirm whether a positive test result is a true positive (i.e. due to the presence of the suspect in the mixture) or, instead, whether it is due to the presence of a relative in the mixture.

2.3. Response to the Homer Method

In response to the Homer *et al.* paper (Homer *et al.*, 2008), a number of other articles have been published to address the implications of participant identification in genome-wide association studies (GWAS). The need to balance research productivity with participant privacy has been emphasised, and different opinions have been expressed regarding how best to manage genomic data (P3G Consortium *et al.*, 2009; Thorisson *et al.*, 2009). One possible solution is to maintain restricted access to genomic data, but to simplify the procedure to apply for access by creating universal researcher IDs. These IDs would potentially allow access to different genomic databases, and would thus prevent the need to apply separately for access to different related databases (as is the situation currently). Another possible solution is to criminalise acts of identifying participants and breaching participant confidentiality in GWAS, for example, in much the same way that breaching patient confidentiality is a criminal offence for doctors. Alternatively, a further solution proposed is to gain informed consent from participants to use genomic data without ever promising full confidentiality (Lunshof *et al.*, 2008).

Another possible solution to countering the problems posed by the Homer method involves developing alternative formats for the reporting of results from

GWAS (Little *et al.*, 2009). For example, it may be possible to avoid the risk of participant identification by publishing only summarised results, such as the distributions of p-values and effect sizes. This issue is addressed in Section 2.9, with the aim of clarifying precisely what information can and cannot be published safely from GWAS.

Although the Homer *et al.* paper has provoked reaction both from governing bodies and prominent researchers in the field of genomics, concerns over the generalisability (or *external validity*) of the Homer method have been raised. In sections 2.2.2 to 2.2.5 we considered some of the problems with the Homer method, such as its reliance on composite hypotheses, and issues regarding the reference population. The reliability of the Homer method in practice is therefore questionable. Recent publications explore the method's performance in more detail, and extensions to the method are also proposed (e.g. in the form of new or adapted test statistics). These publications are now outlined.

2.3.1 Previous Findings

In the context of GWAS data, Braun *et al.* (Braun *et al.*, 2009) specifically examine the Homer method's reliance on three assumptions: (1) that the mixture, reference group and individual of interest are all from the same underlying population; (2) the reference group and the mixture are similarly sized; and (3) the SNPs used in the test are independent. Using data from the International HapMap Project (International HapMap Consortium, 2003), they find that the specificity of the test (i.e. the proportion of individuals who are not in the studies that are *correctly* inferred as absent from the studies) dramatically diminishes when the test individuals are of different ethnicity to the individuals

within hypothetical case-control studies (i.e. which represent the mixture and the reference group). This finding is supported by Sampson and Zhao (Sampson *et al.*, 2009) (see later), who show that the type I error rate of the Homer method can exceed the power when the reference group and the mixture are ancestrally unmatched. The implications of the assumption of “co-ancestry” are investigated further in Section 2.8.

Braun *et al.* also show that the specificity of the test is again reduced when the SNPs used are in LD. As predicted in Section 2.2.5, the test statistic has greater variance when correlated SNPs are used and, hence, the number of false-positive results increases. Although Braun *et al.* do not explicitly test the influence of having different sizes for reference groups and mixtures, they argue that uneven sample sizes also affect how the test performs. For instance, the allele frequencies in groups with greater sample sizes will be more representative of the allele frequencies in the underlying population and, hence, under the null hypothesis, an individual will be “closer”, on average, to the larger group than to the smaller sized group (assuming that the assumption of co-ancestry applies).

In addition to testing the assumptions of the Homer method, Braun *et al.* examine the accuracy of the Homer method in practice by calculating the positive predictive value (PPV). The PPV is a Bayesian measure that provides the probability that a positive test result is a “true positive”. Hence, for the Homer method, the PPV is the probability that an individual inferred as present within the mixture is actually present in the mixture. In addition to the false-positive and the true-positive rates, the PPV depends on the prior probability

that a test individual is in the mixture. For the particular specificity (i.e. 1-false positive) and sensitivity (i.e. true positive) rates deduced for the Homer method, Braun *et al.* report that a PPV of 90% requires a prior probability of the test individual being in the mixture of at least 0.66. Thus, the prior “suspicion” that a test individual is in the mixture must be at least 0.66 in order to be 90% certain that an individual identified within the mixture is actually in the mixture. Consequently, Braun *et al.* conclude that the Homer method would rarely be of use in practice due to the limited specificity of the test.

2.3.2 Extensions of the Homer method

As stated earlier, various extensions of the Homer method have aimed to develop a better and more reliable test for identifying participants from genomic mixtures using SNP allele frequencies. This section outlines these alternative approaches.

2.3.2.1 Jacobs *et al.*

A test proposed by Jacobs *et al.* (Jacobs *et al.*, 2009) specifically addresses the case-control situation [i.e. application (c) in Section 2.2.3], and, hence, is based on a two-sample problem. The statistic proposed by Jacobs *et al.*, T_{geno} , compares genotype frequencies in the reference group and in the mixture for the genotypes of the individual of interest. This contrasts to the Homer method, which compares the *allele* frequencies. Where $X_{j, g}$ and $Y_{j, g}$ represent the genotype frequencies for genotype g of the j^{th} SNP in the reference group and the mixture respectively, and where Z_j denotes the genotype for the individual of interest, the Jacobs *et al.* distance metric is:

$$d = \sum_{j=1}^S \log X_{j,Z_j} - \log Y_{j,Z_j}.$$

A two-sided T-test is proposed to discriminate between the hypotheses that the individual of interest is in neither of the two test groups (H_0); in the reference group (H_1); or in the mixture (H_2). The authors also claim that the test can also assess a fourth hypothesis that the individual is in both groups (H_3); however, although the distance metric has specific expectation under H_3 , it is not clear how the test could discriminate between this hypothesis and other hypotheses *post-hoc*.

Consistent with the Homer method, the power of the test statistic T_{geno} increases with the number of SNPs used, and is negatively correlated with the size of the mixture and the genotyping error rate. In a simulation study, test groups of 1,000 individuals require between 50,000 and 70,000 SNPs to achieve a power close to 100% at $p\text{-val} < 10^{-6}$. The number of SNPs required for this approach, thus, seems similar to the Homer method (although exact figures for the Homer method are not provided). When the reference group is increased to 10,000 individuals, however, a similar power (97%) to infer presence in the mixture is achieved with only 25,000 SNPs (again at 10^{-6}). Using case-control GWAS data, Jacobs *et al.* fit a series of logistic regression models to determine the strength of association between a number of SNPs and a phenotype of interest. The SNPs are ranked by strength of association, and the sensitivity and specificity to infer presence in the case group are derived for various numbers of top associated SNPs. For a mixture of 1,000 cases (with 1,000 controls as the reference), the 1,000 top associated SNPs yield a sensitivity of 43% at a specificity of 95%, but the sensitivity approaches

zero at a specificity of 99.9% (i.e. a 1 in a 1,000 chance of obtaining a false-positive result). Again for a mixture of 1,000 cases (and with 1,000 controls), the 5,000 top associated SNPs yield a sensitivity of 90% at a specificity of 95%, and a sensitivity of 41% at a specificity of 99.9%. For a mixture of 5,000 cases, the 20,000 top associated SNPs yield a sensitivity of approx. 65% at a specificity of 95%, and a sensitivity of approx. 10% at a specificity of 99.9%. The use of top associated SNPs therefore increases the power of the test compared with using randomly chosen SNPs, but many thousand top associated SNPs are still required to reliably infer presence in a mixture of 1,000 or more individuals.

In the Jacobs *et al.* paper, all empirical results are based on equal sample sizes for the two test groups used (i.e. the reference group and the mixture, or the case and control groups of a study). Although the authors do not explicitly state whether their test statistic is influenced by differing sample sizes (as has been highlighted by Braun *et al.* (Braun *et al.*, 2009) as important), they assume that the expectation of the statistic is approximately equal to zero under the null hypothesis. Unequal sample sizes would be likely to cause small deviations of the test statistic from zero under the null, but the nature of these deviations, in effect, would be random (i.e. with mean zero). The test statistic, thus, would not be systematically biased towards one group when the sample sizes are unequal (as is the case with the Homer method). Furthermore, the estimated variance of the test statistic would not be biased if the two sample sizes differ because it is a function of the two sample sizes. Hence, this test statistic *appears* to be robust to unequal sample sizes, although this was not specifically documented.

2.3.2.2 Visscher *et al.*

Visscher *et al.* (Visscher *et al.*, 2009) demonstrate two alternative approaches to testing for presence in a mixture of DNA: a likelihood ratio approach, and a linear regression approach. The likelihood-ratio statistic compares the probability that the individual of interest is in the mixture with the probability that he/she is “out” of the mixture. For the linear regression approach, an outcome is derived by subtracting the allele frequencies in the reference group from the scaled genotypes for an individual of interest, and this outcome is regressed on an explanatory variable derived by subtracting the allele frequency in the reference group from the allele frequency in the mixture. If the individual is in the mixture, the regression coefficient has an expected value of 1, but if he/she is not, it has an expected value of 0. This linear regression approach is distinct from all the other approaches that have been proposed for this problem, and I examine this method in more depth in sections 2.5 to 2.7.

Both test statistics proposed by Visscher *et al.* have similar properties to one another and to the statistics described previously. For instance, the tests have power proportional to the number of SNPs used and inversely proportional to the size of the mixture; allele frequencies (i.e. rare or common) have little influence on the tests except at the extremes; and the tests have greater power using known allele frequencies for a reference population than using estimated allele frequencies from a reference group. In their simulations, Visscher *et al.* compare the linear regression test statistic to the Homer method using simulated sets of independent SNPs. The linear regression statistic generally has greater power than the Homer method, with the exception of when the size

of the reference group is smaller than the size of the mixture. In this situation, the Homer method has the greater power – but this comes at the expense of a sometimes drastically elevated type I error. The linear regression approach consistently yields acceptable type I error rates – even when the sample sizes of the two test groups are unequal. Visscher *et al.* state that a limitation of their test statistics is that the reference group, mixture, and the individual of interest are all assumed to be from the same population. In fact, this has been assumed by all the approaches considered so far. The next approach to be described, however, attempts to overcome the reliance on this assumption.

2.3.2.3 Sampson & Zhao

In an attempt to counter some of the problems potentially posed by the assumption of co-ancestry, Sampson and Zhao (Sampson *et al.*, 2009) propose a new test statistic that requires only that the reference group is matched to the individual of interest. This is a clear advantage over the other methods for a forensic application of the test because, in practice, the ancestry (or *ethnicity*) of the mixture may be unknown, whereas the ethnicity of the individual of interest would be available. The Sampson and Zhao test statistic is derived by calculating a distance metric for the individual of interest and, in addition, for each individual in the reference group. This distance metric is similar to that derived for the Homer method, but, for each SNP, instead of taking *absolute* differences between a particular test individual's scaled genotype and the allele frequency in the reference group and between the test individual's scaled genotype and the allele frequency in the mixture, these differences are *squared*

(it is, thus, an L_2 distance metric). The Sampson and Zhao statistic can be expressed as follows.

For the j^{th} SNP, the individual of interest has scaled genotype Y_{ij} (=0, 0.5 or 1), the allele frequency in the reference group is γ_{Rij} (=0 to 1) and the allele frequency in the mixture is γ_{Mij} (=0 to 1). The distance metric, $D_{L_2,j}$, is then:

Equation 16

$$D_{0,L_2,j} = (Y_{ij} - \gamma_{Rj})^2 - (Y_{ij} - \gamma_{Mj})^2$$

In addition to deriving $D_{0,L_2,j}$, a distance metric is also derived for each of the N_R individuals in the reference group in the same way. Hence, for the i^{th} individual in the reference group, the distance metric is $D_{i,L_2,j}$ (calculated as for $D_{0,L_2,j}$ in Equation 16 above). Subsequently, this metric is averaged over each member of the reference group:

$$\bar{D}_{L_2,j} = \sum_{i=1}^{N_R} D_{i,L_2,j}$$

For each SNP, $\bar{D}_{L_2,j}$, is then subtracted from the distance metric for the individual of interest, $D_{0L_2,j}$:

$$D_{L_2,j} = D_{0,L_2,j} - \bar{D}_{L_2,j}$$

Finally, as with the Homer method, $D_{L_2,j}$ is averaged over all SNPs, and a one-sample t-test is performed. Where D_{L_2} is

$$D_{L_2} = \frac{\sum_{j=1}^S D_{L_2,j}}{S},$$

and where:

$$\hat{\sigma}_D = \frac{\sum_j^s (D_{L_2,j} - D_{L_2})^2}{s} + \frac{\sum \sum_{j_1 \neq j_2}^s (D_{L_2,j_1} - D_{L_2})(D_{L_2,j_2} - D_{L_2})}{s},$$

the T-statistic, T_{L_2} , is:

Equation 17

$$T_{L_2} = \frac{D_{L_2} * \sqrt{s}}{\sqrt{\hat{\sigma}_D}}.$$

In simulations, this test statistic demonstrates a higher power than the Homer method when the ancestries of the reference group, mixture and individual of interest are the same. Although it is assumed that this new test statistic is robust to ancestral differences between the reference group and the mixture, no empirical tests are reported to show this. Furthermore, because the test requires individual level data for the reference group, it is unsuitable for GWAS applications where such data are not always available.

2.3.2.4 Clayton

An alternative approach to test for presence within a mixture of DNA has been proposed by Clayton (Clayton, 2010) using a Bayesian framework. A Bayes factor, $P(\text{Data}|H_1)/P(\text{Data}|H_0)$, is initially derived to test for presence in a mixture when the allele frequencies for a reference population are known. Assuming a bivariate normal distribution for the distance between an individual of interest and the mixture and the distance between the individual and the reference group, for a particular SNP, the individual has genotype x , the mixture has allele frequency \bar{x} , and the reference population has allele frequency μ . The proposed Bayes factor is thus as follows:

$$\log \text{Bayes Factor} = \frac{s}{2} \log \frac{n}{n-1} - \frac{1}{2} \left\{ \frac{n}{n-1} (x - \bar{x})^T (x - \bar{x}) - (x - \mu)^T (x - \mu) \right\},$$

where s is the number of independent SNPs, and n is the sample size of the mixture.

Extensions of the above Bayes factor are proposed for various situations, such as when the allele frequencies in the reference group are estimated (i.e. by use of a reference group) and when they are completely unknown. If the allele frequencies in the reference population are known, the relationship between the number of SNPs required and the size of the mixture is approximately proportional to n ; if no information is available for a reference population, the number of SNPs required is approximately proportional to n^2 ; and if estimates of the allele frequencies for the reference population are available (i.e. by use of a reference group with sample size m), the number of SNPs required (assuming large m) is proportional to $\frac{1}{2} \left(\frac{1}{n} - \frac{1}{n+m} \right)$.

Clayton also considers the situations where the ancestry of the two groups differs, and where there is LD. Where the mixture and the reference group differ in ancestry, Clayton suggests adjusting the Bayes factor for Wright's F_{ST} statistic (Wright, 1968), which measures the degree of genetic divergence between populations. This adjustment requires the ancestry of the two test groups to be known (as well as the corresponding F_{ST} value) and, in effect, reduces the power of the test. If the specified F_{ST} value is inaccurate, it can also perturb the results. To account for linkage disequilibrium (LD), the inverse correlation matrix, Σ^{-1} , needs to be estimated from a secondary dataset. However, Σ^{-1} will be very large and sparse (i.e. it has dimensions equal to the

number of SNPs used in the test), so Clayton suggests using least angle regressions (LAR) (Efron *et al.*, 2004) to estimate its diagonal and near-diagonal elements only.

The method is illustrated in a real data example using a reference group of 1,455 British individuals and a mixture of 145 Scottish individuals, and using 4,743 SNPs on chromosome 20. Hence, the SNPs used are likely to be in LD, and there may be differences in ancestry between the two test groups. In general, the Bayes factor adequately discriminates between individuals present in the mixture and individuals absent from the mixture in these data. After adjusting for LD, the variance of the Bayes factor is reduced and, hence, fewer false-positive results are obtained. In these data it is shown that no adjustment for ancestry is actually required; nevertheless, an adjustment for a relatively extreme F_{ST} value merely leads to marginally more conservative test results.

2.3.2.5 Sankararaman *et al.*

A further test statistic – proposed by Sankararaman *et al.* (Sankararaman *et al.*, 2009) – aims to clarify the maximum number of SNPs that can be published safely for any given study. This test uses a likelihood ratio to infer presence within a genomic mixture by specifying the joint distribution of the genotype for an individual of interest, x , and the allele frequency in the mixture, \hat{p} , under the null and alternative hypotheses. Sankararaman *et al.* claim that their test provides an upper bound on the maximum power achievable by any test by way of the Neyman-Pearson lemma, which guarantees that no test can achieve greater power than the likelihood ratio test. Furthermore, they claim that the powers achieved in their simulation studies are optimal given that *known* allele

frequencies for a reference population are used in contrast to *estimated* population allele frequencies – which lead to a loss of efficiency.

In a real data illustration of the method, an *approximate* likelihood ratio statistic (i.e. which uses estimated allele frequencies for the reference population) yields greater power than the Homer method. As the number of SNPs is increased, the power of the approximate likelihood ratio statistic also appears to converge to the optimal power obtained for the *exact* likelihood ratio statistic (i.e. which uses *known* population frequencies).

The relationship between the power, $1-\beta$; the level of type I error, α ; the number of SNPs, s ; and the sample size of the mixture, n , is described in the following formula:

$$z_{1-\beta} + z_{\alpha} \approx \sqrt{(s/n)},$$

where z refers to respective values from the normal distribution. This provides the theoretical power of the exact likelihood ratio statistic given a fixed sample size and number of SNPs, and a specified type I error rate.

Based on the above formula, Sankararaman *et al.*, have developed an online tool called “SecureGenome”, which calculates the number of top-ranked SNPs that can be published safely at different power and type I error thresholds (see Section 2.9.1). Although this tool is potentially useful for quantifying the number of SNPs that can be released safely in ideal settings in which no model assumptions are breached, it may potentially be more restrictive than is really necessary. Subsequent sections investigate how these types of methods

perform in different realistic scenarios, where their assumptions do not necessarily hold.

2.4. Testing the Original Homer Method

In order to understand the real risks implied by the methods described by Homer *et al.* (Homer *et al.*, 2008), the behaviour of the tests needs to be quantified in terms of their reliance on the underlying assumptions, which are often violated in real data. This section demonstrates how the Homer method performs with respect to some of the conditions and characteristics described in Section 2.2.2. Simulation studies are conducted using different numbers of independent SNPs, and with various reference group and mixture sizes. However, I assume no violation of either of the two key assumptions (i.e. co-ancestry and independent observations) here. Note that a maximum of 50,000 SNPs are used to illustrate the method, as this is the approximate number of independent SNPs in the human genome (Visscher *et al.*, 2009). The focus in these simulation studies is on applying the method in the context of GWAS – rather than in the forensic application.

2.4.1 Simulation Method

These simulation studies investigate how the Homer method performs with various numbers of SNPs, and with various mixture and reference group sizes. Figure 14 below illustrates the basic premise of the simulation method. Allele frequencies for the underlying reference population are simulated first, before genotypes for individuals in the mixture, the reference group, and in neither group are generated from these.

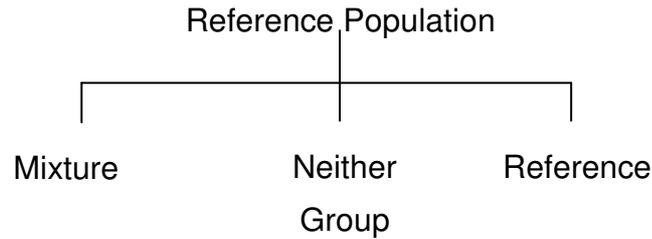


Figure 14: Illustration showing the relationship between the individuals in the mixture, the reference group, and the test individuals who are in neither group. The individual of interest can be in any of these groups.

For the j^{th} SNP, the minor allele frequency (MAF) in the reference population, Pop_j , is generated randomly from a uniform distribution with parameters 0.05 and 0.5 ($j = 1, \dots, s$). This follows the real distribution of MAFs on chromosome 1 in the Utah residents with central European ancestry (CEU individuals) in the International HapMap Project (International HapMap Consortium, 2003), but assumes independence. Individuals in the reference group and the mixture, as well as test individuals in neither group, are then simulated as follows. For the i^{th} individual ($i = 1, \dots, N_k$) in the k^{th} group ($k = 1$ for the reference group; $k = 2$ for the mixture; $k = 3$ otherwise), the j^{th} genotype ($j = 1, \dots, s$), y_{ijk} , is generated randomly from a binomial distribution with $p = Pop_j$ and $n = 2$ ($y_{ijk} = 0, 1$ or 2 copies of the minor allele). Each genotype is then divided by two and, hence, converted to a proportion (or *observed allele frequency*), Y_{ijk} ($= 0, 0.5$ or 1). MAFs for the reference group, G_j , and for the mixture, M_j , are derived by taking the mean genotype (in proportion form) of each SNP within the appropriate group (i.e. $G_j = Y_{.j1}$; $M_j = Y_{.j2}$). Under the alternative hypothesis, each individual in the mixture is tested in turn for presence in the mixture. Under the null hypothesis, the N_3 individuals in neither group are tested in turn for presence in

the mixture. Hence, the test is performed N_2 times under the alternative hypothesis and N_3 times under the null hypothesis (in each run of the simulation), and the hypotheses tested in these scenarios are “not in the mixture” (H_0) Vs “in the mixture” (H_1). Note that the additional situation where an individual is in the reference group is also tested later (see sections 2.6 and 2.7).

Scenario 1 simulates various numbers of independent SNPs (s) with reference groups of 1500 individuals ($N_1=1500$) and mixtures of 500 individuals ($N_2 = 500$). Since it is unclear in the original paper, I do not assume that equal sample sizes are required here. Scenario 2 simulates 50,000 independent SNPs for various N_1 and N_2 values. In both scenarios and in each simulation run, the number of individuals simulated in neither group (N_3) is equal to N_2 . Each individual in the mixture and each individual in neither group is tested in turn for presence in the mixture, using the test hypotheses stated in (b) of Section 2.2.3 (i.e. I only test for presence in a particular GWAS cohort here). Hence, a null hypothesis that the individual is neither in the mixture nor the reference group is tested against an alternative hypothesis that (s)he is in the mixture. Various numbers of simulation runs are performed depending on the values of N_2 and N_3 , ensuring that a minimum of 20,000 tests are performed in each setting. In each simulation run, a new set of MAFs for the reference population, Pop_j , are generated and, hence, new sets of individuals (i.e. who are in the mixture, the reference group, and in neither group) are also simulated.

2.4.2 Scenario 1: Number of SNPs

In this scenario, Monte Carlo estimates of the sensitivity (i.e. the proportion of individuals in the mixture correctly inferred as present in the mixture) and 1-specificity (i.e. the proportion of individuals in neither group incorrectly inferred as present in the mixture) are obtained for the Homer method. Figure 15 overleaf illustrates the results in Receiving Operating Characteristic (ROC) curves, using various numbers of SNPs (s) when $N_1 = 1500$ and $N_2 = 500$. A reference group of 1,500 individuals is chosen here because it represents the approximate size of the Wellcome Trust Case-Control Consortium (WTCCC) element of the British 1958 Birth Cohort (Power *et al.*, 2006), which is a representative sample of the UK population. A mixture size of 500 individuals is chosen for computational convenience, and because this represents an approximate minimum size for cohorts that contribute to consortia of genome-wide association study (GWAS).

As can be seen, when fewer than 10,000 SNPs are used to test for presence in the mixture, a reasonable sensitivity (e.g. 80% or greater) is only obtained at low specificities (e.g. less than 90%). Use of 10,000 or 20,000 independent SNPs yields reasonable sensitivity at a specificity of up to 99%, but the sensitivity diminishes rapidly at more stringent specificities. A consistently high sensitivity close to 100% is obtained with the use of 50,000 independent SNPs, and this remains the case even at 99.999% specificity. In this hypothetical scenario in which none of its model assumptions are breached, these findings concur with findings reported elsewhere (Homer *et al.*, 2008; Jacobs *et al.*, 2009; Sankararaman *et al.*, 2009), and indicate that the Homer method has

sufficient power to reliably infer presence in a mixture of DNA so long as a sufficient number of SNPs is available for use in the test. Use of an insufficient number of SNPs, however, may result in inaccurate inferences regarding whether or not a particular individual of interest is present in any given mixture. Note that Section 2.7 considers further scenarios in which the model assumptions no longer hold.

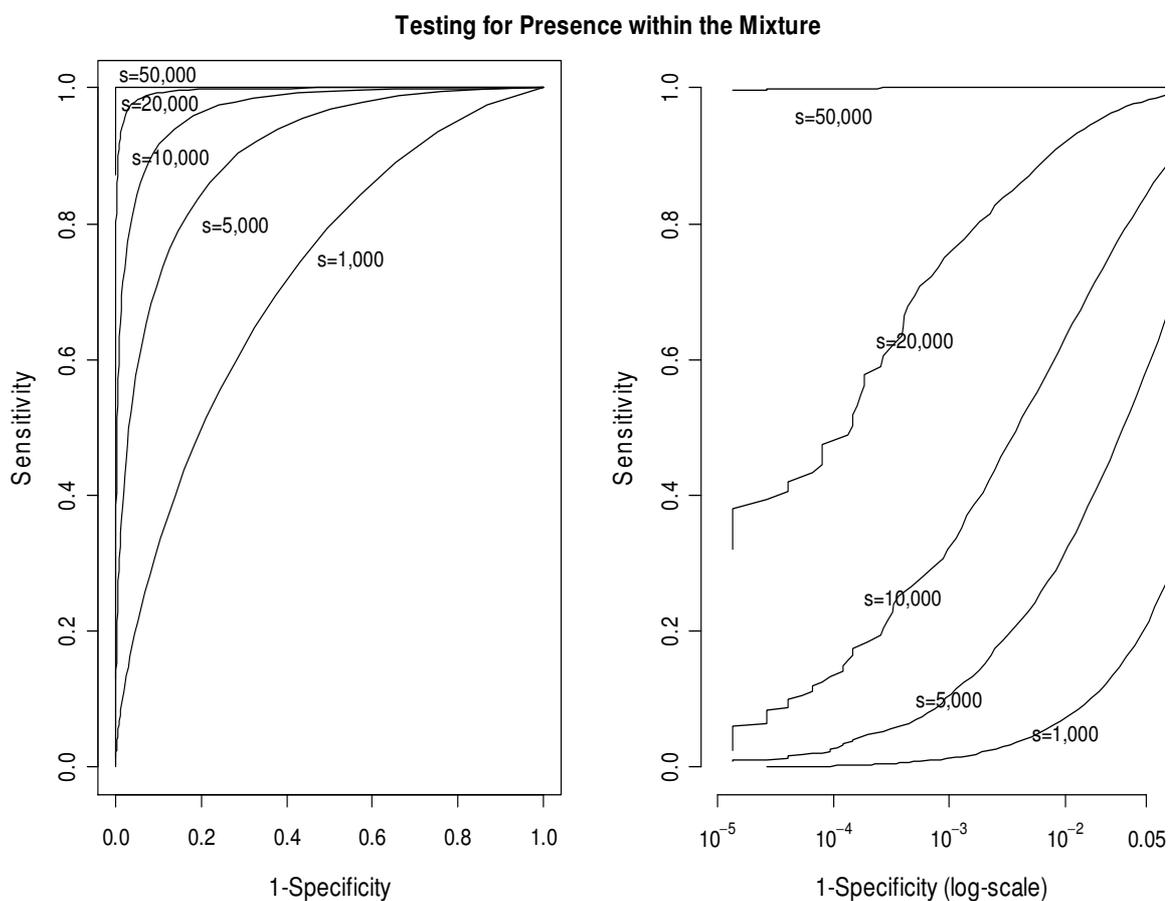


Figure 15: ROC curve showing the sensitivity and 1-specificity of the Homer method when various numbers of SNPs (s) are used. A fixed reference group size (N_1) of 1,500 and a fixed mixture size (N_2) of 500 are used.

2.4.3 Scenario 2: Reference Group and Mixture Size

In this scenario, the effect of different reference group and mixture sizes (N_1 and N_2 respectively) is illustrated using $s = 50,000$ SNPs. Rather than showing the sensitivity and specificity, results are presented by plotting histograms of the Homer test statistic, $T(Y_i)$. In all plots, the putative null distribution is shown in a dotted blue line.

The influence of the size of the reference group is first illustrated (see Figure 16). Against a mixture of size 500, reference group sizes of $N_1 = 90, 500, 1500$, and infinity are tested – noting that a reference group of infinite size is equivalent to knowing the population MAFs, which Homer *et al.* stipulate using in the original paper (Homer *et al.*, 2008). Under the alternative hypothesis it can be seen that the size of the reference group has no influence on the test statistic: the distribution of $T(Y_i)$ for individuals in the mixture is the same for each value of N_1 tested (mean ≈ -6), and is distinct from both the putative null [i.e. which is approximately $N(0,1^2)$ for large numbers of observations] and from the actual distribution of $T(Y_i)$ for individuals in neither group. Importantly, however, under the null hypothesis, the distribution of the test statistic relates directly to the size of the reference group (N_1) in relation to the size of the mixture (N_2). For instance, when N_1 is greater than N_2 , the distribution of the test statistic is to the right of the putative null distribution, and when N_1 is less than N_2 , the distribution of $T(Y_i)$ is to the left of the expected distribution. Only when N_1 equals N_2 does the distribution of $T(Y_i)$ approximate to the putative null. Hence, this confirms the statements of Braun *et al.* (Braun *et al.*, 2009). The implications of having a mismatched empirical and putative null distribution

has been described in Section 2.2.2, and will be discussed again in Section 2.4.4.

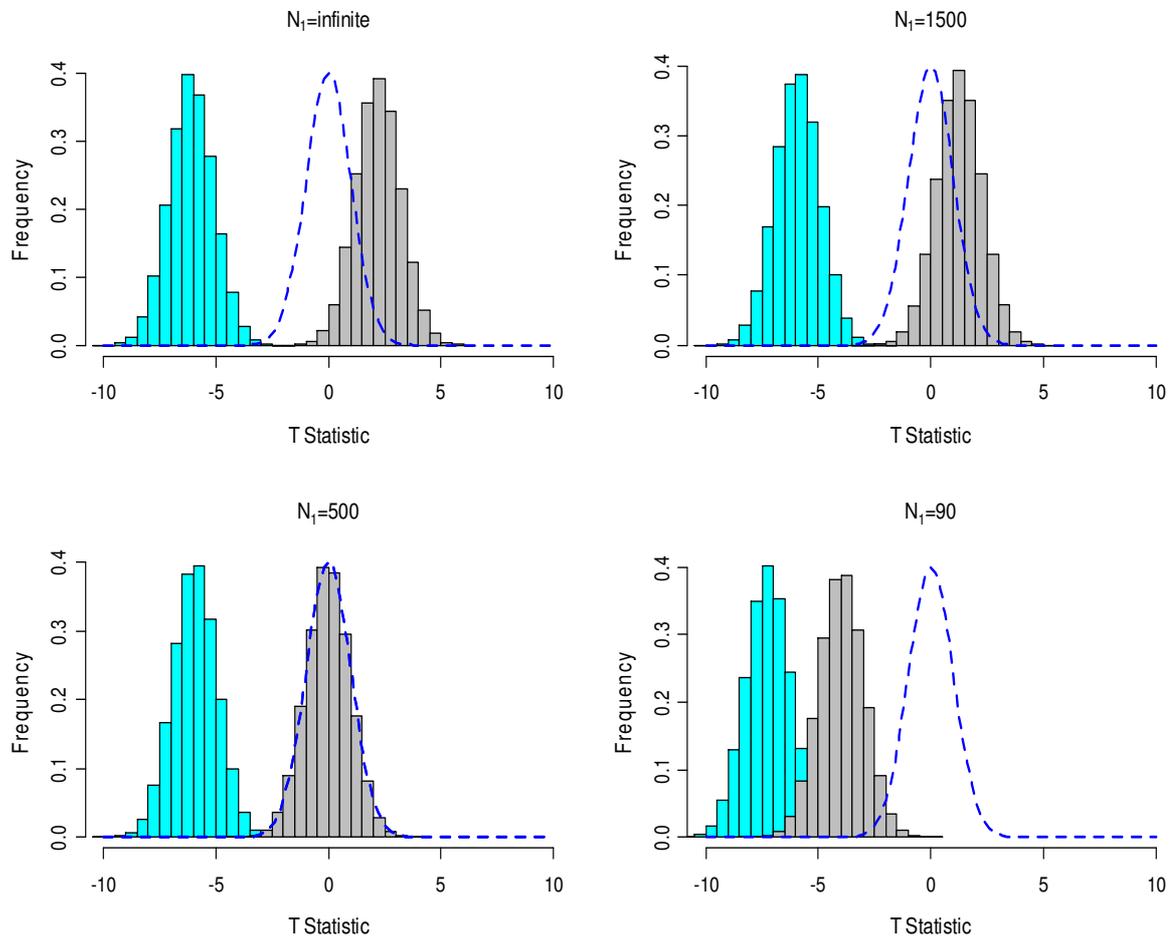


Figure 16: Histograms of the test statistic under the alternative (blue) and null hypotheses (grey). Various reference group sizes (N_1) are tested when the mixture size (N_2) is fixed at 500 and $s = 50,000$ SNPs are used. The navy dotted line shows the expected distribution of T-statistic values under the null hypothesis.

To illustrate how the size of the mixture influences the performance of the Homer method, histograms of the test statistic, $T(Y_i)$, are again provided (see Figure 17 overleaf). In these simulations, the number of SNPs is again fixed at $s = 50,000$, and the reference group size (N_1) is fixed at 1,500 individuals. Mixture sizes of $N_2 = 90, 200, 500$ and 1500 individuals are tested.

As can be seen, under the null hypothesis, the distribution of the test statistic, $T(Y_i)$, is again dependent on N_1 relative to N_2 . When N_1 and N_2 are both 1,500 the distribution of the test statistic conforms to the theoretical (i.e. putative) null distribution, but when N_2 is smaller than N_1 the distribution of $T(Y_i)$ is shifted to the right of the putative null. Under the alternative hypothesis, the influence of the mixture size is clear. For large mixtures the distribution of the test statistic is located closer to the null distribution, while for smaller mixtures the test statistic is located away from the null distribution. Hence, there is greater sensitivity (or *power*) to identify individuals in smaller mixtures. This is logical, as it should be more difficult to identify an individual in a large group than in a small group.

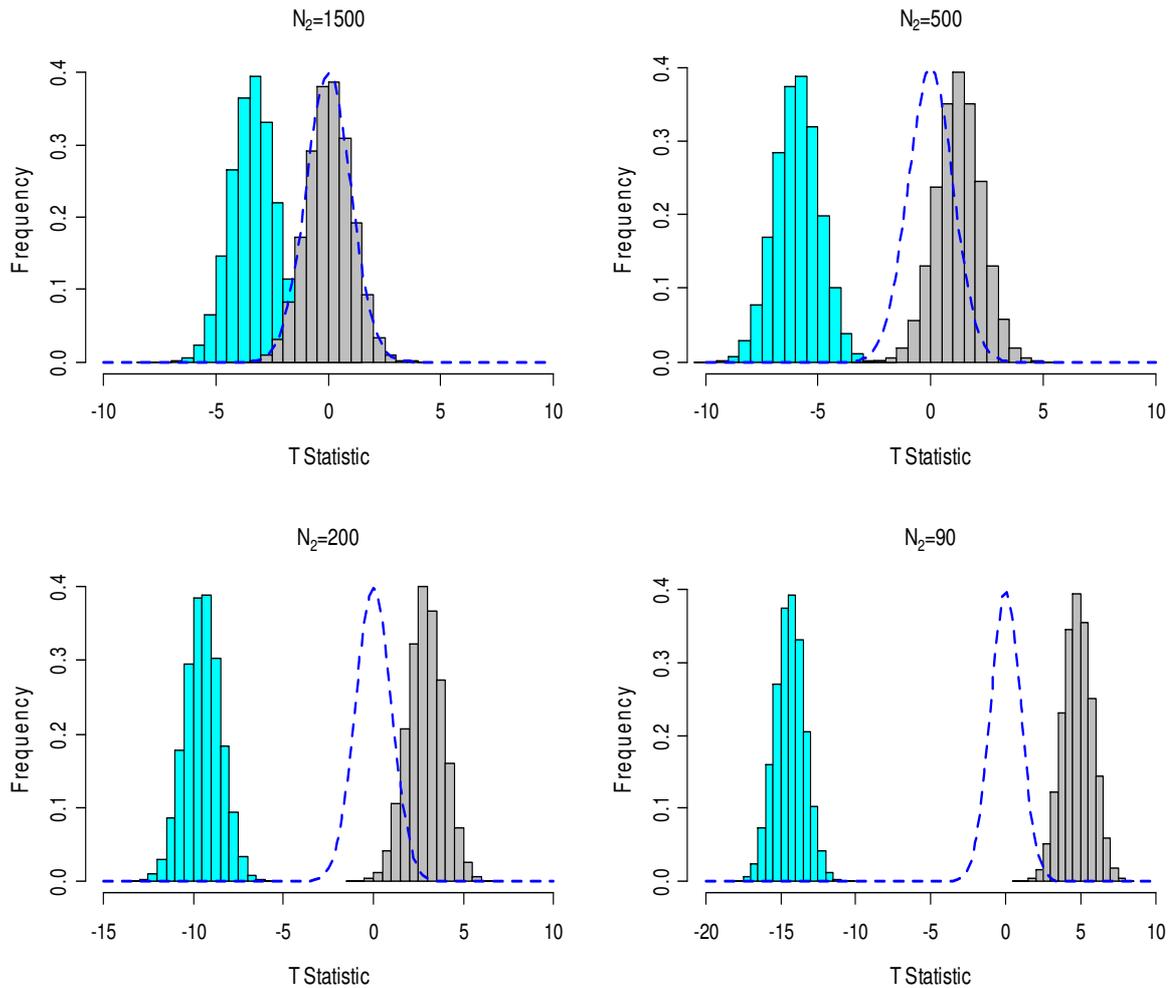


Figure 17: Histograms of the test statistic under the alternative (blue) and null hypotheses (grey). Various mixture sizes (N_2) are tested when the reference group size (N_1) is fixed at 1500 and $s = 50,000$ SNPs are used. The navy dotted line shows the expected distribution of T-statistic values under the null hypothesis.

2.4.4 Discussion of the Results

As shown in Section 2.4.2, the power (i.e. sensitivity) of the Homer method is increased as greater numbers of SNPs are incorporated into the test. Similarly, there is a greater power to infer the presence of an individual in mixtures with smaller sample sizes. These findings are both to be expected, and support the results reported by Homer *et al.* (Homer *et al.*, 2008). In Section 2.4.3, different

sample sizes for the reference group and the mixture are simulated, and it is shown that unless the sample sizes of the two groups are equal (i.e. $N_1 = N_2$) the distribution of the test statistic under the null hypothesis is moved away from the assumed, t-distribution (i.e. which can simply be considered as the standard normal distribution, due to the large number of observations required). This finding can be understood by considering the allele frequencies in the reference group and in the mixture as estimates of the allele frequencies in the underlying, reference population. When the sample sizes of the two groups are equal, the allele frequency estimates are equally precise, and an individual who is not in either group should, on average, be equidistant to both groups (assuming that the assumption of co-ancestry applies). However, when one group is larger than the other, the allele frequencies in the larger group should be more representative of those in the reference population and, as a result, an individual will be “closer” to the larger group than to the smaller group under the null hypothesis.

As described in Section 2.2.2.2, if the reference population allele frequencies are known, a one-sided test of the null hypothesis that mean $D(Y_{ij}) = 0$ is conservative because $D(Y_{ij})$ is actually expected to be less than or equal to zero (i.e. the assumed null value reflects the “worst-case” scenario). Where a reference *group* is used, if the reference group has a greater sample size than the mixture, this principle again applies; the test will be conservative because mean $D(Y_{ij})$ is expected to be ≤ 0 under the null hypothesis. When the sample size of the reference group is smaller than that of the mixture, however, individuals will tend to be closer to the mixture than to the reference group under the null hypothesis, and an increase in type I error would be expected

(i.e. where individuals are incorrectly inferred as present in the mixture). This is an important observation, which was first pointed out by Braun *et al.* (Braun *et al.*, 2009) but which has largely been ignored elsewhere in the literature. Although there is no doubt that having a larger sample size for the mixture than the reference group is problematic, it could be argued that the reverse situation (i.e. where the reference group is greater than the mixture) is desirable because it leads to a conservative test. In many instances – such as when a sufficient power is available – a conservative test could be seen as desirable because it reduces the chance of obtaining a false-positive result. However, as is quoted by Fisher in Section 2.2.2.2, the inexact specification of the null distribution in this situation is a problem because it prevents the accurate derivation of p-values. The Homer method yields biased p-values if the two test groups have unequal sample sizes and, hence, any conventional inference based on such a test will be flawed.

The problem of unequal group sizes also has implications for an application of the Homer method in case-control GWAS data. In this scenario, no reference group is required because the case and control groups can be tested directly against one another. If the sample sizes of the two groups differ the test will be biased towards inferring presence in the larger of the two groups and, as such, it is apparent that the Homer method itself would be unsuitable for use in this context. Thus, as the main focus of this work is on the GWAS applications of the test, it now seems sensible to focus on one of the alternative approaches to the Homer method instead of the Homer method itself.

As has been mentioned above, the effect of unequal sample sizes has largely been ignored in the other papers that propose extensions to the Homer method, and it is not obvious how most of the alternative approaches would be affected by unequal sample sizes. Nevertheless, the paper by Visscher *et al.* (Visscher *et al.*, 2009) includes results for simulation studies testing various samples sizes for the two test groups, and their linear regression approach yields the correct level of type I error in all scenarios. This approach, thus, seems robust to differing sample sizes. Noting that Braun *et al.* (Braun *et al.*, 2009) suggest that the Homer method is of little use in practice due to the high prior probability required that an individual is in the mixture, a further strength of the Visscher *et al.* test is that it seems to outperform the Homer method in most simulation scenarios. The Visscher *et al.* approach thus appears more suitable for the current problem than the Homer method. From here, I therefore focus on assessing the capabilities of the Visscher *et al.* (Visscher *et al.*, 2009) test in preference to the Homer method. In keeping with the initial aims of this work, the subsequent sections focus on understanding the true implications of these methods with respect to their potential threat to participant confidentiality in GWAS.

2.5. Visscher *et al.* Linear Regression

As described in Section 2.3.2.2, Visscher *et al.* have proposed two alternatives to the Homer method: a likelihood ratio approach and a linear regression approach. As will be discussed later, an advantage of these approaches in

comparison to the other approaches described in Section 2.3.2 is that the null hypotheses they test each has a fully specifiable distribution. Hence, the Visscher *et al.* approaches should, in principle, yield the correct level of type I error regardless of the sample sizes of the groups compared in the test (i.e. the reference group and the mixture, or the case and control groups of a study). It is predominantly for this reason that, from here, I consider the Visscher *et al.* approaches in favour of any of the other approaches proposed to identify individuals within genomic mixtures. Note, however, that the approach proposed by Clayton (Clayton, 2010) also appears useful, because it can also handle differences in the sample sizes between the mixture and the reference group. The framework proposed by Visscher *et al.* is more straightforward than the Clayton framework, however, and, as we will see in Section 2.7, it allows for some straightforward extensions that improve its performance in realistic situations.

2.5.1 Overview

Visscher *et al.* state that there is a strong and consistent relationship between their likelihood ratio and linear regression approaches. For instance, the log-likelihood ratio statistic they propose is approximately equal to half of the difference between the linear regression approach's test statistic under the null and alternative hypotheses. It therefore only seems necessary to focus on one of their two approaches here. In this section I thus provide an overview of the linear regression approach only. Note that some alternative notation will now be introduced, and that this new notation will be used from here on in preference to the Homer *et al.* notation used previously in this chapter.

For the j 'th SNP ($j = 1, \dots, s$), the true, underlying population allele frequency is p_j ; the observed allele frequency in the mixture is q_j ; and the individual of interest has genotype y_{ij} , which is in proportion form (= 0, 0.5 or 1 representing 0, 1 or 2 copies of the minor allele respectively). Where a reference group is used to estimate p_j , the observed allele frequency in the reference group is r_j . The mixture consists of N_{mix} individuals, and the reference group, where used, consists of N_{ref} individuals.

The linear regression approach is based on a regression of Y_{ij} on X_j , where Y_{ij} and X_j are deviations of the population frequency p_j from y_{ij} and from q_j respectively:

$$Y_{ij} = b.X_j + \epsilon_{ij},$$

where $Y_{ij} = y_{ij} - p_j$, and $X_j = q_j - p_j$; b is the regression coefficient; and ϵ_j is a normally distributed error term. For known population allele frequencies, p_j , the least squares estimate of b – assuming that the slope of the line is forced through the origin (i.e. there is no intercept term) – is:

$$\hat{b} = \frac{\sum_{j=1}^s (y_j - p_j)(q_j - p_j)}{\sum_{j=1}^s (q_j - p_j)^2}.$$

As is shown in Appendix B.2, the expectation of b is 1 if the individual of interest is in the mixture, and 0 if he/she is not. If the individual of interest is in the mixture, b has variance: $Var(b/in) = (N_{mix} - 1)/s$; otherwise, b has variance: $Var(b/out) = N_{mix} / s$.

As discussed in Section 2.2.2.1, in practice, the population frequencies p_j are not known, and from here I therefore focus on the two sample problem (e.g. in which a reference group is used). Visscher *et al.* express an alternative version of the test for the two sample situation, which is based on the estimated population frequencies, \hat{p}_j . Rather than directly using the allele frequencies in the reference group, r_j , as estimates of \hat{p}_j , Visscher *et al.* recommend using a pooled average of the allele frequency in the mixture, q_j , and the allele frequency in the reference group, r_j , as an estimate of \hat{p}_j . A combined sample of the N_{mix} and N_{ref} individuals in the mixture and in the reference group is thus used to specify \hat{p}_j :

Equation 18

$$\hat{p}_j = \frac{\sum_{i=1}^{N_{mix}+N_{ref}} y_{ij}}{N_{mix}+N_{ref}}$$

In this situation, p_j is replaced in the regression by \hat{p}_j . Hence, the regression is again Y_{ij} on X_j , but Y_{ij} is now $(y_{ij} - \hat{p}_j)$ and X_j is now $(q_j - \hat{p}_j)$. Based on the same principle as before, the estimate of the regression coefficient for this situation is:

$$\hat{b} = \frac{\sum_{j=1}^S (y_j - \hat{p}_j)(q_j - \hat{p}_j)}{\sum_{j=1}^S (q_j - \hat{p}_j)^2}$$

If an individual is in the mixture, \hat{b} again has an expectation of 1. Visscher *et al.* state that if an individual is not in the mixture, \hat{b} has an expectation of 0. However, as is shown in Appendix 2, this is true only for individuals who are neither in the mixture nor in the reference group; if an individual is in the

reference group itself, \hat{b} actually has an expectation of $-N_{mix}/N_{ref}$. This has implications to the hypothesis testing procedure to be used with this approach. As discussed in Section 2.2.3, the only appropriate formulation of a test for a two sample setting (i.e. where an individual could be in either of two test groups or in neither of the two groups) is based on a two-tailed hypothesis. Visscher *et al.* have, however, proposed two one-tailed tests (outlined in Appendix 2), which both seem problematic. Their first test compares a null hypothesis that $b = 0$ (i.e. “not in mixture”) against an alternative hypothesis that $b > 0$ (i.e. “in mixture”). As noted earlier, the problem with this formulation of the test hypotheses is that it ignores the possibility that an individual could also be in the reference group [in which case $E(b) < 0$]. Their second test compares the null hypothesis that an individual is in the mixture ($b = 1$) against an alternative hypothesis that he/she is not ($b < 1$). Although this formulation of the test is based on a null hypothesis with an exact distribution, it does not seem particularly useful in practice. For instance, an underpowered analysis could easily lead to the fallacious conclusion that an individual is in the mixture when he/she is not, because the null hypothesis represents the outcome of interest. Furthermore, assuming that an individual is in the mixture under the null hypothesis seems an unclear and unsatisfactory way of testing for presence in the mixture.

For these reasons, rather than using the chi-square tests stipulated by Visscher *et al.* (which are outlined in Appendix B.2), a Z-test is instead to be adopted to illustrate the method in this chapter:

$$Z = \frac{X - \mu_0}{\sigma} = \frac{\hat{b}}{\sqrt{\text{var}(b|\text{out})}} \sim N(0, 1^2),$$

where $\text{Var}(b|\text{out}) = [N_{\text{mix}}/s]^*[(N_{\text{mix}} + N_{\text{ref}})/N_{\text{ref}}]$, and where “out” implies that an individual is in neither of the two test groups.

If the null hypothesis (*out*) is rejected in the above, a further discriminatory analysis is required to ascertain which of the two groups the individual is in. Given that \hat{b} *must* either be significantly greater than zero or significantly less than zero for the null hypothesis to be rejected, an informal discriminatory analysis might involve simply checking \hat{b} for its sign (i.e. positive or negative) \hat{b} . This approach to performing a discriminatory analysis is examined in Section 2.6.

2.5.2 Assumptions and Practical Implications

As with the other identification approaches, the Visscher *et al.* linear regression approach assumes co-ancestry between the mixture, the reference group, and the individual of interest (see Section 2.2.4). The approach also assumes independent observations.

In addition to the assumptions of co-ancestry and independent observations, the Visscher *et al.* linear regression approach also relies on the standard assumptions of any linear regression. It thus assumes that the error terms are independent and identically distributed following a normal distribution with mean zero. The assumption that the error terms have constant variance regardless of

the fitted value is known as the assumption of *homoscedasticity*, or homogeneity of variance.

As has already been mentioned, an advantage of the Visscher *et al.* approach over some of the other methods that have been proposed to infer presence within a mixture is that it is expected to be robust to differences in the sample sizes of the two groups compared in the test. In a reported simulation study (Visscher *et al.*, 2009), the linear regression approach yields the correct level of type I error for any sample size combination of the reference group and mixture. When compared with the Homer method (Homer *et al.*, 2008), the linear regression approach also yields a greater power in most scenarios. However, if the sample size of the mixture is smaller than the size of the reference group, this test yields a lower power than the Homer method, i.e. because the Homer test has increased power in this situation at the expense of an increased type I error rate.

As to be expected, the linear regression approach requires greater numbers of SNPs to infer the presence of an individual within larger mixtures because the power of the test is increased with the use of more SNPs. For very large reference groups, Visscher *et al.* define the relationship between the number of SNPs required, the sample size of the mixture, and the power/type I error rate of the test as $s/N_{mix} = (z_{\alpha} + z_{1-\beta})^2$, i.e. where α is the significance level and $1-\beta$ is the power. Note that this matches the power of the exact likelihood ratio statistic proposed by Sankararaman *et al.* (Sankararaman *et al.*, 2009). If the reference group and the mixture have equal sample sizes, however, the

Visscher *et al.* statistic requires twice this number of SNPs (i.e. relative to the size of the mixture) to achieve the same power, i.e. $(z_{\alpha} + z_{1-\beta})^2 = 2*s/N_{mix}$.

Although the simulation studies reported by Visscher *et al.* (Visscher *et al.*, 2009) demonstrate that the approach is potentially useful in hypothetical situations in which none of the model assumptions is infringed, more work is required to clarify how well the method performs in more realistic scenarios. In the following section I therefore test the Visscher *et al.* linear regression approach in a number of settings – using both simulated and real data – to determine how effective it really is in practice.

2.6. Testing the Visscher *et al.* Method: Simulation Studies

Before testing the linear regression approach in scenarios where its model assumptions are breached, this section illustrates how the method performs in simulation studies in which all individuals are simulated from the same population, and in which all the SNPs used in the test are independent. This section, thus, aims to replicate the results reported by Visscher *et al.*. These simulation studies test the method in the context of a case-control GWAS, i.e. where an individual could be in either the case or the control group, or in neither group. The focus of these analyses, thus, is on the two-sample application of the method. Hence, where the Visscher *et al.* method is outlined in Section 2.5.1 in terms of a mixture and a reference group, these groups should now be thought of as the case and the control groups of a study.

These simulations test the capability of the linear regression method to accurately infer the presence of an individual within a study as a whole (using a two-tailed hypothesis test). When the null hypothesis is rejected, a discriminatory analysis is performed based solely on the sign of the regression coefficient, \hat{b} , to ascertain whether an individual of interest is a case or a control. Initially, there are no systematic differences between individuals in the case group, the control group, and neither group; all individuals simulated are thus sampled into the test groups randomly (Scenario 1). Subsequently, disease status for a hypothetical disease is simulated based on genotypes for various numbers of “causal SNPs”, and individuals are sampled into the case or control arms of a study according to disease status (Scenario 2). Note that Scenario 1 is the equivalent of having zero causal SNPs in Scenario 2.

2.6.1 Scenario One: Random Sampling

Cases and controls are initially simulated using the same simulation method as reported in Section 2.4. Population minor allele frequencies, p_j , are simulated from a uniform distribution with limits 0.05 and 0.5 ($j = 1, \dots, s$). From an identification point of view, rare alleles with frequency less than 0.05 could be useful, but due to the known problems with genotyping rare SNPs in practice, I do not assume use of SNPs with a rare allele frequency (this approach has also been used elsewhere, e.g. (Homer *et al.*, 2008; Jacobs *et al.*, 2009; Sampson *et al.*, 2009; Visscher *et al.*, 2009)). Genotypes for N_{ca} individuals in the case group; N_{con} individuals in the control group; and $N_{neither}$ individuals in neither group are then simulated from two calls to the binomial distribution (with $p = p_j$). Each genotype is then divided by two and converted to a proportion of the total

possible number of copies of the minor allele, i.e. any individual can possess up to two copies of a particular minor allele (with the exception of those on the sex chromosomes, which I do not assume use of here at all), so the genotypes 0, 1 or 2 are represented by 0, 0.5 or 1 in proportion format. For each SNP, the minor allele frequency (MAF) in the case group, q_j , is derived by calculating the mean genotype (in proportion form) of the N_{ca} cases, and an estimate of the population MAF, \hat{p}_j , is derived by taking a pooled estimate of the allele frequencies in the case and control groups (see Equation 18). In each simulation run, each individual in the case group, the control group, and in neither group is tested in turn for presence in the study using the linear regression approach. Fifty thousand SNPs are simulated in each run, and 100 runs are performed for each combination of N_{ca} and N_{con} . In each simulation run, 1,000 individuals who are not in the study are tested in addition to the individuals in the case and control groups. Results are presented in Table 18 below.

Group	N	Mean (\hat{b})	Var(\hat{b})	Rejections of H_0 :		Discriminatory Analysis (% correct)
				P<0.05	P<10 ⁻⁵	
Case	500	1	0.02	1	0.9938	100
Control	500	-1	0.02	1	0.9944	100
Neither	1000	0.0008	0.02	0.0628	0.00001	-
Case	1000	1	0.06	0.9780	0.3745	100
Control	500	-2	0.06	1	0.9998	100
Neither	1000	0.0016	0.06	0.0638	0.00002	-
Case	500	1	0.015	1	0.9998	100
Control	1000	-0.5	0.015	0.9780	0.3745	100
Neither	1000	-0.0008	0.015	0.0638	0.00002	-
Case	1000	1	0.04	0.9980	0.7098	100
Control	1000	-1	0.04	0.9981	0.7090	100
Neither	1000	-0.0002	0.04	0.06223	0.00004	-

Table 18: Results for 100 runs of Scenario One. Every individual in each group is tested in turn for presence in the study. The proportion of rejections of H_0 represents the power to infer presence in the study for cases and controls, and it represents type I error for individuals in neither group. If the null hypothesis is rejected (at the 5% level of significance), a discriminatory analysis is conducted to ascertain which group the individual is in. Discriminatory analyses are not reported for individuals in neither group who are incorrectly inferred as present in the study.

As can be seen in Table 18, estimates of the regression coefficient b are virtually unbiased, on average, for any combination of N_{ca} and N_{con} , and the type I error rates are all close to the nominal level (they do, however, seem consistently larger than 0.05; I consider this issue further in Section 2.7.2). At the 5% level of significance, the approach has a near 100% power for any sample size combination. At the $p < 10^{-5}$ level of significance (which is a more rigorous threshold that has also been used elsewhere (Homer *et al.*, 2008; Braun *et al.*, 2009; Jacobs *et al.*, 2009)), the power again nears 100% when the sample sizes are both 500 and when a group of 500 individuals is tested against a group 1000 individuals. When both groups have a sample size of 1000 individuals, the power at $p < 10^{-5}$ is around 70%. This reduction in power is

to be expected because, naturally, there will be less power to infer presence in a larger group. When one group is smaller than the other, power to infer the presence of an individual in the larger group is impaired (e.g. power to infer presence in a group of 1000 individuals when the other group consists of 500 individuals is approx. 37%). These results show that, for a given number of SNPs, the sample sizes of the two groups are key in determining the power of the test. Furthermore, they show that the ordering of the two groups (i.e. which allele frequencies are to be taken as q_j , and which contribute only to \hat{p}_j) does not affect the outcome of the test. Where the test correctly implies that an individual is in a study, the discriminatory analysis, which is based on the sign of \hat{b} , has a perfect success rate for correctly inferring case or control status. Hence, the decision to adopt a two-tailed test for the problem at hand seems appropriate.

These results support the simulations reported by Visscher *et al.*, and show that the method performs well in this hypothetical situation where none of its assumptions are explicitly breached.

2.6.2 Scenario Two: Sampling by Disease Status

In this scenario, case and control subjects are sampled into a study according to disease status for a simulated disease. The disease is randomly generated from a logistic regression model in which various numbers of causal SNPs influence the probability of contracting the disease. Individuals with the disease are ascertained into the case group, and individuals without the disease are ascertained into the control group. Two additional groups of individuals who are not in the study are also simulated: one for each disease state. The individuals

within these two groups are also tested for presence in the study (under the null hypothesis).

Population minor allele frequencies (MAFs) for a set of c independent causal SNPs are first simulated. For the j 'th causal SNP ($j = 1, \dots, c$), the population MAF, is generated from a uniform distribution with parameters 0.05 and 0.5. Genotypes for each causal SNP are then simulated for a large population of N individuals, where the j 'th genotype for the i 'th individual ($i = 1, \dots, N$), y_{ij} , is randomly generated from a binomial distribution with two trials with probability equal to the population MAF ($y_{ij} = 0, 1$ or 2 copies of the minor allele). Next, an effect for each causal SNP, β_j , is randomly generated. The simulated SNP effects are log-odds ratios, which, for the j 'th causal SNP, represent the increase in risk of having the disease of interest per copy of the minor allele. β_j is generated randomly from a normal distribution with mean 0 and variance 0.4^2 and, hence, approximately 95% of the odds-ratios are between 0.44 and 2.23.

For the purpose of deriving the linear predictor, LP_i , all genotypes for the causal SNPs are centred by subtracting the expected genotype for each SNP (i.e. which is equal to twice the population MAF of the corresponding causal SNP). For the i 'th individual, LP_i is derived using an additive genetic model, by multiplying the j 'th genotype by the j 'th SNP effect, and summing these terms over all c causal SNPs:

$$LP_i = \beta_0 + \beta_1(y_{i1} - \bar{y}_{.1}) + \dots + \beta_{1c}(y_{ic} - \bar{y}_{.c})$$

where β_0 is an intercept term (the magnitudes of which are presented in Table 18:). Next, a probability of disease, d_i , is derived for each individual using the

inverse-logistic (or *expit*) transformation, $\exp(LP_i)/[1+\exp(LP_i)]$, before disease status, D_i , is simulated by taking a random draw from a Bernoulli distribution with $p = d_i$ ($D_i = 1$ if individual has the disease of interest; $D_i = 0$ otherwise). For the values of β_0, \dots, β_c simulated, approximately 5-10% of the individuals in the population of size N are simulated with the disease of interest.

Four groups of individuals are now randomly sampled from the population by disease status. A group of n_1 controls represents the control group in a case-control GWAS study; n_2 cases represent the case group in the same study; n_3 controls represent a group of control individuals who are not in this study; and n_4 cases represent a group of case individuals who are not in the study. All other individuals from the population (i.e. who are not sampled into one of these four groups) are now discarded. Next, population MAFs for a set of $(s - c)$ independent, *non-causal* SNPs, p_j , are simulated in the usual way. The j 'th non-causal SNP ($j = c+1, \dots, s$), p_j is thus randomly generated from a uniform distribution with parameters 0.05 and 0.5. For each individual sampled into one of the four test groups, a genotype for each non-causal SNP is then simulated – again in the usual way. For the i 'th individual [$i = 1, \dots, (n_1 + n_2 + n_3 + n_4)$], the genotype for the j 'th non-causal SNP ($j = c+1, \dots, s$) is generated randomly from a binomial distribution with $p = p_j$ and $n = 2$. All genotypes [i.e. both for the c causal SNPs and for the $(s-c)$ non-causal SNPs] are now converted to allele frequencies (i.e. between 0 and 1). Each genotype, thus, is divided by two and converted from 0, 1 or 2 to a 0, 0.5 or 1 respectively. For the j 'th SNP, the mean genotype in the case group is taken as q_j , and the mean genotype in the case and control groups combined is taken as \hat{p}_j .

In each simulation run, every individual within each of the four test groups is tested in turn for presence in the study. Fifty thousand independent SNPs are simulated in total in each run, and each group consists of 500 individuals. Simulation characteristics for the present scenario are summarised in Table 19 below.

Property	Value(s)
Number of causal SNPs (c)	50; 250; 500; 1000.
Total number of SNPs (s)	50,000
Population allele frequencies	~Uniform (0.05, 0.5)
Intercept term (β_0)	-3.5; -6; -8; -11.
SNP effects (β_i)	~Normal (0, 0.4 ²)
Population size (N)	20,000
Number of controls (n_1)	500
Number of cases (n_2)	500
Number of test controls not in study (n_3)	500
Number of test cases not in study (n_4)	500

Table 19: Simulation characteristics for each simulated case-control GWAS.

Results for this scenario are shown in Table 20 below.

Group	No. Causal SNPs (c)	Mean (\hat{b})	Var(\hat{b})	Rejections of H_0 :		Discriminatory Analysis (% correct)
				P<0.05	P<10 ⁻⁵	
Case	50	1	0.02	1	0.9943	100
Control		-1	0.02	1	0.9950	100
Not in study - Case		0.0097	0.02	0.0615	0.00002	-
Not in study - Control		-0.0106	0.02	0.0626	0.00002	-
Case	250	1	0.02	1	0.9946	100
Control		-1	0.02	1	0.9941	100
Not in study - Case		0.0185	0.02	0.0617	0.00002	-
Not in study - Control		-0.0182	0.02	0.0628	0.00002	-
Case	500	1	0.02	1	0.9945	100
Control		-1	0.02	1	0.9950	100
Not in study - Case		0.0198	0.02	0.0600	0.00002	-
Not in study - Control		-0.0212	0.02	0.0620	0	-
Case	1000	1	0.02	1	0.9953	100
Control		-1	0.02	1	0.9942	100
Not in study - Case		0.0212	0.02	0.0610	0.00004	-
Not in study - Control		-0.0218	0.02	0.0632	0.00004	-

Table 20: Results for 100 runs of Scenario Two. Every individual in each group is tested in turn for presence in the study. The number of rejections of H_0 represents the power to infer presence in the study for cases and controls, and it represents type I error for individuals not in the study. If the null hypothesis is rejected (at the 5% level of significance), a discriminatory analysis is conducted to ascertain which group the individual is in. Discriminatory analyses are not reported for individuals in neither group, i.e. because these individuals are *incorrectly* inferred as present in the study.

Regardless of the number of causal SNPs generated, the method yields approximately the correct level of type I error. A consistently high power is also obtained, and the discriminatory analysis again has a perfect success rate. These findings support the results shown in the previous scenario and,

furthermore, demonstrate that the Visscher *et al.* method is capable of accurately inferring presence in or absence from a case-control GWAS even when the individuals within the two arms of a study are systematically different. Note, however, that despite the genotypic differences simulated here between individuals with and without the disease, the assumption of co-ancestry still applies (as well as the assumption of independent observations). Note also that although the type I error rates are approximately correct here, as in Scenario 1, they are consistently slightly elevated.

In addition to the results reported here, I also performed a set of more extreme simulation studies to test how the method performs when the cases and controls are more drastically different. Applying the same simulation method as described above, even when up to 100% of the SNPs were designated as causal SNPs, and when the standard deviation of the log-odds of contracting the disease per copy of each minor allele was increased to 1, approximately the correct levels of type I error were retained. Hence, these results suggest that, when simulated in this way, no matter how different two test groups are the method will perform appropriately. Note, however, that the assumption of co-ancestry was always upheld in these simulations. The effects of a breach in co-ancestry are explored in Section 2.8.

Interestingly, although previous work claims that top-ranked SNPs, i.e. SNPs associated with a particular cohort and ordered by strength of association, provide greater power for participant identification (Jacobs *et al.*, 2009; Sankararaman *et al.*, 2009), no evidence of this has been found here. For instance, the power in this scenario does not seem to increase when greater

numbers of causal SNPs are simulated and, similarly, the powers yielded in this scenario are approximately equal to the power yielded in Scenario 1 for the case-control GWAS with 500 participants in each arm. Top-associated SNPs are likely to be more informative in these tests than SNPs with allele frequencies that do not differ significantly between two groups because, typically, they should contribute more to the test statistic. Further testing is therefore required to confirm whether this principle also applies to the Visscher *et al.* approach. As the power in these simulated studies consistently approaches 100% (even at $p < 10^{-5}$), it is difficult to notice whether or not the power truly differs between scenarios.

2.7. Real Data Illustrations & Extensions

We have now seen that the Visscher *et al.* approach performs well in simulated data where neither of the two key assumptions of these methods (i.e. co-ancestry and independent observations) is breached. In practice, however, there is no guarantee that these assumptions will be upheld. For instance, as outlined in Section 2.2.4, real data are likely to be correlated due to linkage disequilibrium (LD) between SNPs, and real studies may be subtly different in terms of ancestry even if they appear well matched. Although simulation studies have been performed to assess the method up until now, the effects both of LD and of differences in ancestry are difficult to simulate realistically. For instance, real LD structures are extremely complex, and the true magnitude of the differences between ancestries is currently unknown. Here, real data is therefore used to illustrate how the method performs in practice, and to examine the implications both of LD and of differences in ancestry. Where possible,

these findings are supported by additional simulation studies, but the results from these simulations will not always be shown.

2.7.1 1958 Birth Cohort

Permission to access genotype data from the 1958 British Birth Cohort (Power *et al.*, 2006) was granted by the WTCCC (2007). The 1958 Birth Cohort (1958BC) consists of 1,504 unrelated participants who were all born in Great Britain in 1958. The region of birth is recorded (i.e. one of twelve UK regions – including Scotland and Wales but excluding Northern Ireland), with similar numbers of participants recruited within each region (number of participants per region ranges between 75 and 160). The genotypes used in these analyses are typed on the Affymetrix 500K chip and called in Chiamo – Oxford format (2007; Marchini *et al.*, 2007). Any genotypes called with a probability of less than 0.9 are omitted from the dataset. Following advice provided in exclusion files that accompany the data, 24 participants and 30,956 SNPs are also omitted completely (in all subjects).

In the following analyses, hypothetical studies are to be simulated using the individual-level data from the 1958BC. As in the simulation studies reported in Section 2.6, a number of simulation runs are to be performed in each scenario in order to obtain Monte Carlo estimates of the power (i.e. the proportion of individuals correctly inferred as present in the study) and type I error rate (i.e. the proportion of individuals incorrectly inferred as present in the study) for the Visscher & Hill linear regression approach. In each simulation run, a set of real individuals is randomly sampled without replacement into the case and the control arms of a hypothetical case-control GWAS, and into a group (or into one

of two groups) of test individuals who are not in the study. Every individual within each of the test groups is then tested in turn for presence in the study – as has been performed in all simulations so far. Since the scaled genotype format is required to perform the Visscher *et al.* method, all genotype data from the 1958BC are converted from 0, 1 or 2 to 0, 0.5 or 1 respectively.

2.7.2 Common Ancestry & No LD

In this initial scenario, I aim to demonstrate how the Visscher *et al.* method performs in real data without explicitly violating the key assumptions of independent observations and co-ancestry. As such, I attempt to avoid LD by using well spaced SNPs from across the genome, and I attempt to avoid differences in ancestry by only selecting individuals from southern UK regions (i.e. from London, southeast, southwest or south England). SNPs are arranged by chromosome and by position, and every 20th SNP on chromosomes 1 to 10 is initially selected for use in the test (giving 14,767 SNPs in total). Given the reported findings (Visscher *et al.*, 2009), this seems a reasonable number of SNPs to infer presence in a mixture of several hundred individuals. There are 461 individuals in total from the southern regions of the UK in the 1958BC data and, on average, each individual has 65 missing genotypes (number of missing genotypes ranges between 13 and 250). In each simulation run, 100 individuals are randomly sampled without replacement into each arm of a hypothetical case-control GWAS, and 100 individuals are randomly sampled into a test group of individuals absent from the study (i.e. these are test individuals under the null hypothesis). The rest of the simulation method is consistent with the notation and the procedure outlined in sections 2.5 and 2.6. For the j^{th} SNP ($j =$

1,...,s), the allele frequency in the case group represents the allele frequency in the mixture and is denoted q_j . The combined allele frequency in the case and the control groups is used as an estimate of the population allele frequency p_j , and is denoted \hat{p}_j . Every individual within each of the test groups is tested in turn for presence in the study. For a given individual of interest, if a genotype is missing at a particular SNP, that SNP is omitted from the present test only (i.e. of his/her presence within the study).

The simulation is performed 1,000 times and Monte Carlo estimates of the power and type I error rate are obtained for the Visscher *et al.* approach. As before, the mean estimate of the regression coefficient is also derived for each test group, and a measure of the accuracy of the discriminatory analysis (outlined in Section 2.6.1) is provided. Results are summarised in Table 21 below.

Group	N	Mean (\hat{b})	Mean [Var(\hat{b})]	Rejections of H_0 :		Discriminatory Analysis (% correct)
				P<0.05	P<10 ⁻⁵	
Case	100	1	0.0152	1	0.9984	100
Control	100	-0.9999	0.0152	1	0.9984	100
Neither	100	0.0002	0.0152	0.1169	0.0005	-

Table 21: Results for SNP spacing of 20, using individuals only from southern UK regions.

As can be seen, the power of the study is near 100% at the 10⁻⁵% level of significance, and the discriminatory analysis has a perfect success rate.

Participants of the hypothetical studies are therefore correctly identified both as present in the studies – and in the correct arm of the studies – with high accuracy. However, for the test group of individuals absent from the studies (i.e. individuals in neither group) the type I error rate is elevated above the expected rate; for instance, approximately 12% of these individuals are incorrectly inferred as present in the studies at the 5% level of significance. Because in Section 2.6 the Visscher *et al.* approach yields approximately the correct level of type I error in simulated data, one or more characteristics of this real dataset must be causing this increase in type I error. As the regression coefficient is, on average, estimated without bias in this scenario, the elevated type I error here is likely to be due to a problem with the variance. To investigate this, a histogram of the Z-test statistic under the null hypothesis is provided in Figure 18 below.

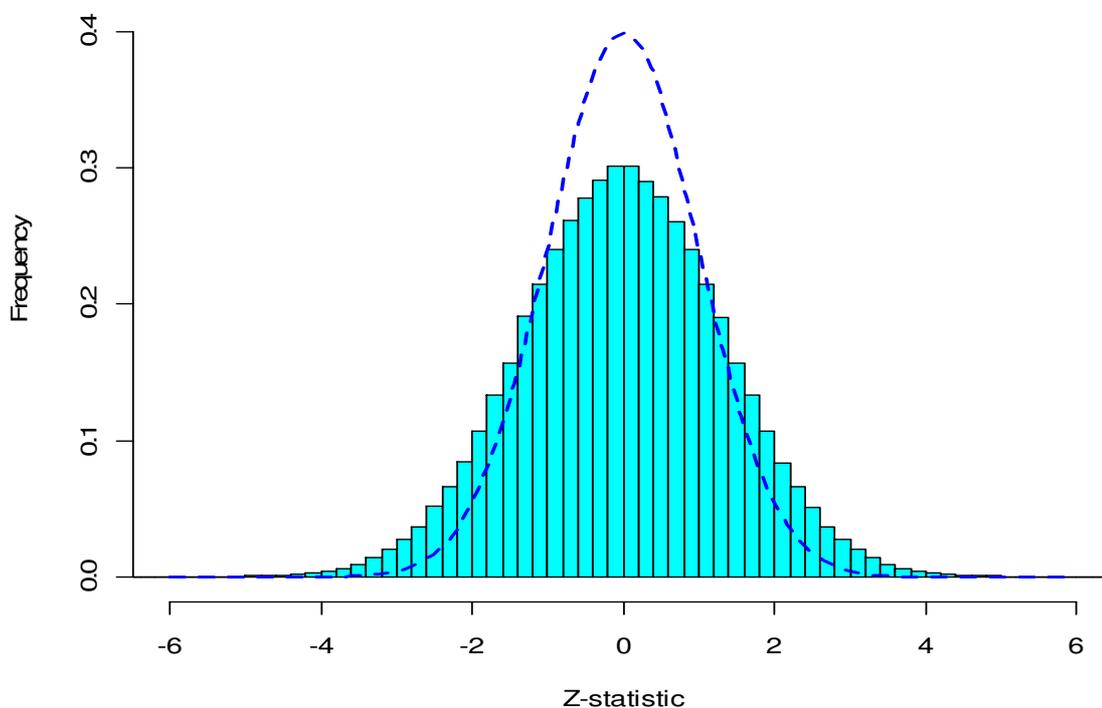
Individuals Absent from Study

Figure 18: Histogram of the z-statistic under the null hypothesis. Results are based on 1,000 runs of a simulation using the individual level data from the 1958BC. All individuals are sampled from southern regions of the UK, and every 20th SNP is utilised from chromosomes 1 to 10. The dotted blue curve shows the expected standard normal distribution of the test-statistic.

Figure 18 shows that the observed z-statistic has greater variance than expected under the putative null distribution. For instance, although the mean of the distribution is correct, the tails of the distribution are wider than the putative distribution, and the spike at the median is shorter. This causes an increase in type I error because a greater number of observations lie in the tails of the distribution than expected – and a greater number of values therefore exceed the pre-specified “critical values”. An obvious cause of this increase in the variance of the test statistic – as explained in Section 2.2.4 – would be LD.

For instance, if some of the SNPs used in the test are in LD, the variance of the test-statistic would be expected to be greater than assumed because correlated observations tend to be less informative than equivalent numbers of independent observations. Although I attempted to avoid LD in the current analysis by spacing the SNPs across the genome, it is possible that some of the selected SNPs lie within regions of the genome with dense LD and, thus, are correlated with each other. I now attempt to confirm whether the increased type I error obtained here is actually due to LD by increasing the SNP spacing and repeating the analyses.

Table 22 below shows results for real data simulations where the SNP spacing is 20, 33 and 100 (spread across chromosomes 1 to 22). With spacing of 100, only 4577 SNPs remain available for use in the test; hence, for consistency between the three analyses, the analyses with SNP spacing of 20 and 33 are limited to using the first 4577 SNPs. As each of these analyses uses fewer SNPs in this scenario compared with the number of SNPs used in earlier scenarios, the powers yielded here are lower (e.g. approx. 0.53 at $p < 10^{-5}$). Nevertheless, the main focus of these analyses is on the type I error rates (i.e. the number of rejections of H_0 for individuals in neither study group). As can be seen, as the SNP spacing is increased the type I error rates decrease closer to the nominal rate. For example, the type I error rate reduces from approx. 0.11 with SNP spacing of 20 to approx. 0.08 with SNP spacing of 100 (at 5% level of sig.). This suggests that LD *is* truly present in the datasets with SNP spacing of 20 and 33, and that LD is causing an increase in the type I error rates. However, even when the SNP spacing is increased to 100, the type I error rate remains considerably greater than the nominal level. Although further

increasing the SNP spacing might allow us to confirm whether or not LD is solely responsible for the elevated type I error rates obtained here, use of a set of SNPs with even greater spacing is impractical because too few SNPs would remain with which to adequately power the test. A previous study has estimated that there are around 55,000 independent SNPs in the human genome, however (Purcell *et al.*, 2009), and SNP spacing of 100 (for those SNPs assayed on the Affymetrix 500K chip) already seems relatively wide. I therefore suspect that the SNPs in this dataset are largely independent. Noting that, even in simulated data, the type I error rates obtained for the linear regression approach are consistently higher than the expected level (e.g. in Section 2.6 the type I error rate is consistently around 0.06 at $p < 0.05$), these findings lead me to suspect that a further characteristic of the data is also affecting the test.

SNP Spacing	Group	Mean (\hat{b})	Mean [Var(\hat{b})]	Rejections of H_0 :		Discriminatory Analysis (% correct)
				P<0.05	P<10 ⁻⁵	
Every 20 th SNP	Case	1	0.0494	0.9802	0.5276	100
	Control	-0.9999	0.0494	0.9809	0.5272	100
	Neither	-0.0002	0.0494	0.11383	0.0004	-
Every 33 rd SNP	Case	1	0.0489	0.9864	0.5356	100
	Control	-0.9999	0.0489	0.9858	0.5382	100
	Neither	-0.0005	0.0489	0.0947	0.0001	-
Every 100 th SNP	Case	1	0.0492	0.9886	0.5321	100
	Control	-1	0.0492	0.9891	0.5318	100
	Neither	-0.0002	0.0492	0.0840	0.0000	-

Table 22: Results for the Visscher et al. linear regression approach applied to the 1958BC data with SNP spacing of 20, 33 and 100. All groups consist of 100 individuals and all individuals are sampled from south UK regions. All analyses use the first 4,577 equally spaced SNPs are selected from chromosomes 1 to 22.

A violation of the co-ancestry assumption is unlikely to be responsible for the elevated type I error rates yielded here because all the participants selected for these analyses are from a similar region of the UK (this conclusion is actually confirmed in Section 2.8). A further, plausible explanation for these increases in type I error, thus, is that the model mis-specifies the variance function. As stated in Section 2.5.2, heteroscedasticity occurs in a linear regression when a model's residuals (or *errors*) do not share a common variance. For example, it can occur when the error terms depend on the values of the explanatory variables, X (i.e. where $x_j = q_j - \hat{p}_j$ in this setting) – or it can be *unsystematic*,

and follow no particular pattern (Martin, 2000). For the data used in these analyses, the errors are actually likely to be heteroscedastic because, assuming Hardy-Weinberg equilibrium, the model outcome includes the genotypes, y_{ij} , which follow a binomial distribution. The variance of the binomial distribution varies with the mean (i.e. the variance function is $p(1-p)/2$ in this case (McCullagh *et al.*, 1991)) and, hence, any assumption of common variance will be violated.

Heteroscedasticity typically biases estimates of the variance (although, under some circumstances, it can also bias estimates of the regression coefficients), and the pattern of results obtained here, thus, seems consistent with this explanation. However, given that the genotypes simulated in Section 2.6 are generated as binomially distributed variables, the results for those studies should also be affected. As the type I error rates reported in Section 2.6 are only marginally greater than the expected level, this issue was not initially recognised. However, these reported error rates are consistently (if only slightly) higher than the expected rate and, in hindsight, they are likely to be a consequence of the mis-specified variance function. Nonetheless, the issue of why the type I error rates shown in Table 22 above are higher than those obtained in Section 2.6 remains to be explained. One possibility is that, even in the dataset with SNP spacing of 100, LD between some of the SNPs remains, and the corresponding correlation causes biased estimates of the variance. Another possibility – which is not mutually exclusive to the previous explanation – relates to the distribution of allele frequencies in the real data (see Figure 19 below). While the simulations in Section 2.6 generate minor allele frequencies (MAFs) from a uniform distribution bounded by 0.05 and 0.5, the analyses

reported in this section use allele frequencies between 0 and 1 – without omitting the SNPs with $MAF < 0.05$. Figure 19 below displays a histogram of the allele frequencies in the 1958BC dataset. As can be seen, there is a sharp peak in the frequency of SNPs with MAFs below 0.05 or above 0.95, and this will impact on any estimates of the variance under assumed homoscedasticity. Because the allele frequencies for these SNPs have substantially lower variance than those for SNPs with greater MAFs (e.g. $var = 0.02375$ Vs 0.12 for a SNP with an MAF of 0.05 Vs 0.4 respectively), datasets skewed towards these values will underestimate the variance and, hence, there will be increases in type I error.

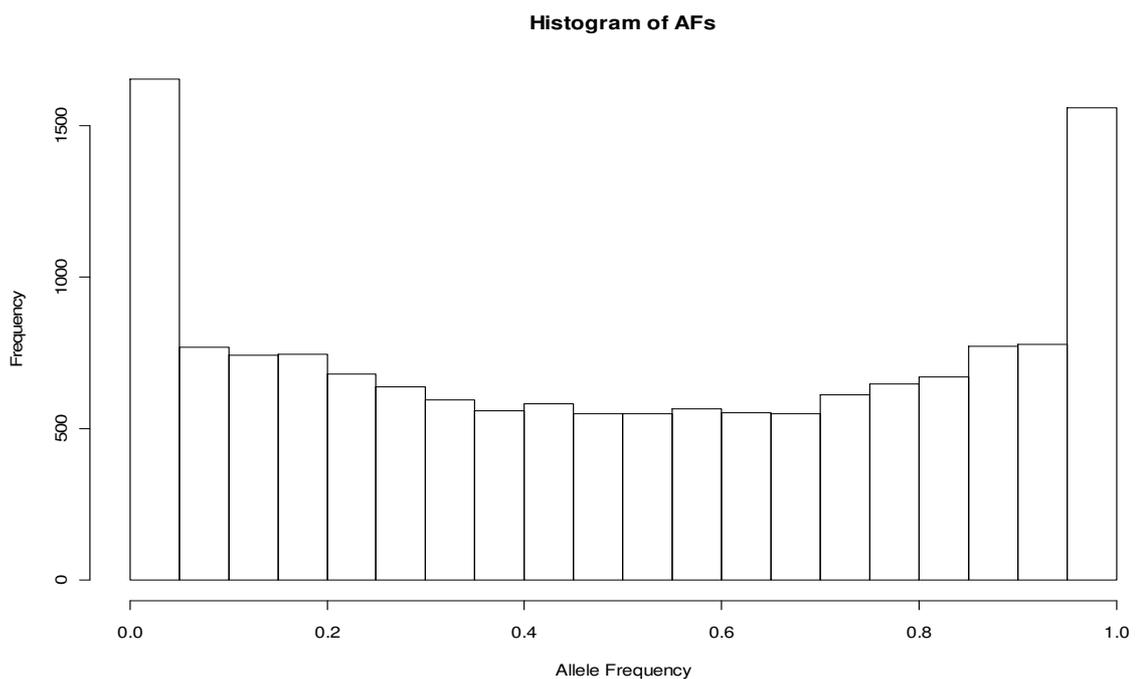


Figure 19: Histogram of allele frequencies (AFs) for every 20th SNP on chromosomes 1-10.

The following section attempts to extend the Visscher *et al.* linear regression approach to better handle the genotype data modelled here. For instance, different approaches to specifying the variance function are proposed with the aim of deriving a model that yields the correct type I error rates in real data.

2.7.3 Modelling the variance

There are potentially numerous ways in which to adjust the Visscher *et al.* linear regression to more appropriately model the variance function, and this section outlines two such approaches. The first approach reformulates the Visscher *et al.* linear regression approach to derive a logistic regression. As such, this approach will correctly model the variance function for binomially distributed data (McCullagh *et al.*, 1991). The second approach uses an independence estimating equation (Liang *et al.*, 1986) – a type of *generalised* estimating equation (GEE) – and is similar to the original approach. GEEs are usually used to allow for correlation without the need to be specific about the correlation structure. I initially use a GEE approach here as a means to derive a *robust* estimate of the standard error of the regression coefficient. Robust standard errors rely on fewer assumptions than conventional standard errors, and provide consistent estimates of parameter standard errors when observations are correlated (Zeger *et al.*, 1986; Burton *et al.*, 1998).

2.7.3.1 Logistic Regression

The original Visscher *et al.* approach regresses Y_{ij} on X_j in a linear model, where, using the same notation as before, $Y_{ij} = y_{ij} - \hat{p}_j$ and $X_j = q_j - \hat{p}_j$. However, the genotypes, y_{ij} , are inherently binomial in nature – taking the values 0, 1 or 2

only (or the equivalent proportions: 0, 0.5 or 1 respectively), and a model in the binomial family therefore particularly suits these data. Generalised linear models with a logistic link – often referred to as *logistic regression* models – are usually applied to binary (or binomial) outcome data, and predict the log odds of an event, Y , as a linear function of the explanatory variables, X (McCullagh *et al.*, 1991). This contrasts with linear models, which predict the change in Y for each unit (or level) change in X . Linear models assume that the variance of each observation is independent of its mean value. On the other hand, generalised linear models allow the variance to vary in a range of different ways. For instance, logistic regression assumes that the variance is a known function of the predicted mean (see formula in Page 167) (Dobson, 2002). Logistic regression, thus, does not assume homoscedasticity, and could be appropriate for the current problem. This section describes a logistic regression approach based on the same principles as the Visscher *et al.* linear regression.

The original Visscher *et al.* outcome, Y_{ij} , has variance that increases with the underlying, population allele frequency, p_j , and the assumption of homoscedasticity, thus, is violated here. In order to correctly model the variance function, Y_{ij} can be converted to the log-odds scale with the aim of fitting a logistic regression to the data.

A log-odds transformation of the original outcome subtracts the *log-odds* of \hat{p}_j from the *log-odds* of y_{ij} [i.e. $\log\left(\frac{y_{ij}}{1-y_{ij}}\right) - \log\left(\frac{\hat{p}_j}{1-\hat{p}_j}\right)$], but this outcome is not compatible with a logistic regression because it is not binomially distributed (only the y_{ij} term is). An outcome that *is* compatible with a logistic regression can most straightforwardly be derived simply by dropping the log-odds of \hat{p}_j

from the left-hand side of the regression (and therefore regressing the log-odds of y_{ij} on X_j). This outcome would depart from the original principles of the original statistic, however. In keeping with these original principles, I wish to maintain the effects of the \hat{p}_j term here (see Section 2.7.3.3 for a discussion of why the \hat{p}_j term is important in these tests and for some results for a model that does not include this information). Hence, rather than dropping the log-odds of \hat{p}_j from the left-hand side of the regression, I instead suggest *offsetting* the log-odds of \hat{p}_j to the right-hand side of the regression. In effect, this simply adds the log-odds of \hat{p}_j to each side of the regression equation. Offset variables can be used to adjust analyses without impacting on the precision of other parameter estimates because they have a gradient fixed at one. Offsetting the log-odds of \hat{p}_j in this situation, thus, allows a logistic regression to be fitted to the observed genotypes, y_{ij} , and maintains the original properties of the Visscher *et al.* regression (i.e. by adjusting for \hat{p}_j in the model). Note, however, that this logistic regression is not equivalent to the original linear regression; it merely follows the same principles as the original approach.

This logistic regression approach predicts the log-odds of having the genotype, y_{ij} (which is fitted as the integers 0, 1 or 2 rather than as their scaled equivalents – see below) given the log-odds of the estimated population allele frequency, \hat{p}_j , and given that the difference between the allele frequency in the mixture, q_j , and \hat{p}_j is X_j (i.e. $X_j = q_j - \hat{p}_j$):

$$\log\left(\frac{y_{ij}}{1-y_{ij}}\right) = \log\left(\frac{\hat{p}_j}{1-\hat{p}_j}\right) + b \cdot X_j; \text{ error} = \text{binomial}.$$

Although logistic regression models often deal with binary outcomes, they can also readily handle binomial outcomes such as the genotype data we have here. For instance, if an outcome is binomial it is merely treated as a series of n independent binary (or *Bernoulli*) observations at a particular level of the covariates, X . For genotype data, n is always 2 and, hence, a genotype of 2 (= 2 copies of the minor allele) is handled as two “successes” at the same level of the covariates, X_j ; a genotype of 1 (= 1 copy of the minor allele) is handled as one success and one “failure” at X_j ; and a genotype of 0 (= 0 copies of the minor allele) implies two failures at X_j . In order to specify the model in R , I use the `glm()` package. This requires that the number of successes (i.e. y_{ij}) and failures (i.e. $2 - y_{ij}$) are specified in separate columns. Alternatively, an equivalent way to specify the model would be to fit the genotype data, y_{ij} , as two separate binary observations, a_{1ij} and a_{2ij} (= 0 or 1 copy of the minor allele) - denoting the two alleles an individual of interest possesses at a particular SNP, j .

2.7.3.2 GEE Independence Model

Generalised estimating equations (GEEs) are often used to derive consistent estimates of regression coefficients and their standard errors in correlated data. They are particularly useful when the nature of the correlation itself is not of primary interest, as they do not require the correlation matrix to be correctly specified. The GEE approach therefore seems particularly attractive for the handling of LD – which I consider in Section 2.7.4. Here, however, I initially assume independence and focus on allowing for heteroscedasticity.

GEEs predict the *marginal* expectation of the observations as a linear function of the covariates, so the regression coefficients from GEE models are therefore interpreted as *population-averaged* effects (Zeger *et al.*, 1986; Zeger *et al.*, 1988). This contrasts with conventional regression methods, which typically model the expectation of the response variable *conditional* upon the covariates. For example, where the regression coefficients from a GEE model might represent a difference in the response between two groups averaged over any other covariates in the model, the regression coefficients in a conventional model represent the effect of changing between different levels of a covariate for a particular individual, i.e. at particular levels of the covariates. Although this difference in the interpretation of the regression coefficients can be important, the two interpretations are actually the same for models with a normally distributed outcome and an identity link. For the application of GEEs I consider here, the outcome proposed by Visscher *et al.* is used throughout (i.e. $y_{ij} - \hat{p}_j$). This will be handled as a normally distributed variable (even though it is constrained between -1 and 1) and, hence, no distinction between the two interpretations of the regression coefficients is required.

GEE models derive estimates of the regression coefficients and the variance-covariance matrix via an iterative process that ensures both are consistent as long as the mean structure is correctly defined. This process involves first fitting a regression model to derive initial estimates of the parameter coefficients and the model residuals. For a GEE extension of the Visscher *et al.* linear regression approach, this regression model is the same as the one outlined in Section 2.5.1, i.e.:

Equation 19

$$y_{ij} - \hat{p}_j = b. (q_j - \hat{p}_j) + \epsilon_{ij}$$

The residuals, ϵ_{ij} , are then used to estimate the correlation parameters (which have pre-specified structure), before the above model is refitted by applying an algorithm that incorporates the estimated correlation coefficients. This procedure is then iterated until the algorithm converges, i.e. when the estimates stabilise (Burton *et al.*, 1998).

Although conventional standard errors can be used with GEEs, these tend to underestimate the variance in the presence of correlation (Dobson, 2002). The *robust* standard error (SE) (White, 1982; Royall, 1986; Williams, 2000) is therefore more commonly preferred. Robust SEs can account for the clustering of data, where, for example, related or correlated observations are ordered together, i.e. in *clusters* (see Section 2.7.4.1).

The robust SE is sometimes known as the “sandwich” estimator because it “sandwiches” the *score* statistic, C , between the inverse of the “information” matrix, J (Hardin *et al.*, 2007). Where y_i denotes the vector of observations for the i^{th} cluster ($i = 1, \dots, N$), X_i denotes the design matrix for the i^{th} cluster, and \hat{V}_i denotes the variance-covariance matrix for the i^{th} cluster, the robust variance, V_S is:

$$V_S = J^{-1} \cdot C \cdot J^{-1},$$

where

$$J = \sum_{i=1}^N X_i^T \hat{V}_i^{-1} X_i$$

and

$$C = \sum_{i=1}^N X_i^T \hat{V}_i^{-1} (y_i - X_i \hat{b}) (y_i - X_i \hat{b}) \hat{V}_i^{-1} X_i.$$

The robust variance, thus, multiplies the components for each cluster separately before summing over all clusters to provide overall estimates of the variances.

Robust SEs converge in probability to their true value (that is, they are *consistent*) even when the specified correlation structure is incorrect and when the correlation differs in different clusters of observations. A GEE approach based on the robust SE, thus, potentially allows correction both for heteroscedasticity and for LD (see Section 2.7.4). Nevertheless, the robust SE can only account for correlation within clusters, i.e. it assumes no between-cluster correlation. Moreover, an incorrect specification of the correlation structure will typically lead to a loss of efficiency.

GEEs are fitted using an iterative procedure that handles each cluster of observations separately before summing over all clusters. Following on from the above notation, the expected vector of observations for the i^{th} cluster is $E(y_i) = \mu_i$; $g(\mu_i) = X_i b$, and D_i is a matrix of the derivatives $\delta \mu_i / \delta b$. The score statistic, U , for Equation 19 is thus:

$$U = \sum_{i=1}^N D_i^T V_i^{-1} (y_i - \mu_i) = 0,$$

where the matrix, V_i is defined as:

$$V_i = A_i^2 R_i A_i^2 \phi.$$

Here, A_i is a diagonal matrix containing the elements $\text{Var}(y_i)$ along the diagonals, R_i is a correlation matrix with pre-specified structure, and ϕ is a constant that allows for overdispersion.

The independence estimating equation described by Liang and Zeger (Liang *et al.*, 1986) is a special case of the GEE approach, which assumes that repeated observations on the same individual (or in the same *cluster* – see Section 2.7.4.1) are independent. The specified correlation matrix for an independence estimating equation model, R_i , is thus the identity matrix. In the case of the data we have here, all the observations in any given test correspond to one particular individual of interest, and this model assumes that each SNP is independent. This model, thus, is identical to the original Visscher *et al.* linear regression, but uses the robust (or *sandwich*) estimator of the variance.

2.7.3.3 Testing the Models

This section tests the approaches to adjusting for heteroscedasticity using the 1958BC data as described in sections 2.7.1 and 2.7.2. The dataset with SNP spacing of 100 is initially used in order to minimise the chances of selecting SNPs that are in LD. Again, only individuals from the southern UK regions are sampled into the three test groups, to avoid breaching the assumption of co-ancestry. In total, a maximum of 4577 SNPs are thus available for use in the test, and a population of 461 individuals with southern UK ancestry is available from which to sample the test groups. As before, in each simulation run a case-control GWAS is sampled consisting of 100 participants in each arm, and a group of 100 individuals absent from the study is also sampled (i.e. to test under the null hypothesis). Again, as before, genotypes for each individual are represented by y_{ij} , the allele frequency in the case group (or *mixture*) is q_j , the underlying, population allele frequency is p_j , and a pooled estimate of the allele frequencies in the case and control groups is \hat{p}_j . Each individual from each of

the three groups is tested in turn for presence in the study, and Monte Carlo estimates of the power and the type I error rates are derived. Four approaches are tested in total here: the original Visscher *et al.* linear regression approach, the logistic regression approach described in Section 2.7.3.1 with the log-odds of \hat{p}_j included as an offset, a logistic regression approach with no offset (i.e. simply the log-odds of y_{ij} regressed on X_j), and the independence estimating equation approach described in Section 2.7.3.2. The statistical package R is used to perform all analyses. The first three approaches are applied using the `glm()` package (using an identity link for the linear regression approach, and using a binomial link for the two logistic regression approaches), and the independence estimating equation approach is applied using the `geepack()` package (specifying an identity link and an independence correlation structure) (Yan *et al.*, 2004; Halekoh *et al.*, 2006).

Results for these analyses are presented in Table 23 below. For all individuals, the original linear regression approach and the independence estimating equation approach yield identical estimates of the regression coefficient, \hat{b} , but, as expected, they differ in terms of the estimates of the variance. For instance, the variances from the original linear regression approach are around 0.05, but the robust variance estimates from the GEE independence model are around 0.06. The GEE independence model, consequently, has a marginally reduced power compared to the linear regression approach (i.e. power \approx 0.980 Vs 0.988 respectively) but, importantly, it also yields a lower level of type I error. Because heteroscedasticity is an issue in the problem we consider here, the GEE independence model thus seems a more appropriate choice of approach than the original Visscher *et al.* linear regression.

Analysis	Group	Mean (\hat{b})	Mean [$\text{Var}(\hat{b})$]	Reject. H_0 (5% level of sig.)	Discriminatory Analysis (% correct)
Linear Regression	Case	1	0.0492	0.9885	100
	Control	-1	0.0492	0.9875	100
	Neither	0.0006	0.0492	0.0837	-
Logistic Regression w/out Offset	Case	4.0955	1.2917	0.9363	100
	Control	-3.9297	1.2909	0.9194	100
	Neither	0.0851	1.2838	0.0746	-
Logistic Regression w/ Offset	Case	5.2277	1.6833	0.9801	100
	Control	-5.2267	1.6830	0.9798	100
	Neither	0.0031	1.6719	0.0528	-
GEE Independence Model	Case	1	0.0606	0.9809	100
	Control	-1	0.0606	0.9800	100
	Neither	0.0006	0.0617	0.0526	-

Table 23: Results for analyses of the 1958BC using SNP spacing of 100 and sampling only individuals with southern UK ancestry. For individuals in the case or control groups the % rejections of H_0 represent power for each approach; for individuals in “neither” group the % rejections of H_0 represents the type I error rate for each approach.

For the two logistic regression approaches, i.e. the model that includes the log-odds of \hat{p}_j as an offset (the “offset” model), and the model without an offset (the “no offset” model), the results differ somewhat. For example, *offset* has almost identical type I error rates and powers to the GEE independence model, while *no offset* yields around 5% lower power than the other approaches and an

elevated type I error rate (type I error ≈ 0.075). Hence, although *offset* performs well, *no offset* does no better than the original linear regression approach despite the fact that both logistic regression approaches correctly model the variance function.

A possible explanation for the discrepancy in performance between the two logistic regression models could be that \hat{p}_j acts as a confounder in this situation. Confounders are variables that are associated both with an outcome and with one or more explanatory variables. The consequences of failing to adjust for confounding variables in a model are well known. For instance, omitted confounders can lead to bias in the estimates of parameter coefficients and, hence, to both type I and type II errors (Negassa *et al.*, 2007). The \hat{p}_j term could be a confounder in this situation because, clearly, it relates strongly to the outcome, y_{ij} , but, also, it relates to – and is included within – the regression term, $X_j (= q_j - \hat{p}_j)$. Although this explanation seems reasonable, results for an additional, linear regression approach – which I have not reported until now – cast some uncertainty on this matter. This additional approach regresses y_{ij} on X_{ij} in a linear regression and, thus, seems conceptually similar to the *no offset* logistic regression. Although *no offset* generally performs worse than *offset* in the above analysis, this additional, linear model actually performs very similarly to the original linear regression. The additional model, for example, has marginally increased estimates of the variance compared to the original model and, hence, slightly reduced power and type I error rates. Consequently, because the type I error rates for the original linear regression are generally above the expected level (see sections 2.6 and 2.7.2), this additional approach

actually yields type I error rates closer to the expected level. Thus, in the logistic model, failing to adjust for the (potential) confounding effect of \hat{p}_j seems to impact on the approach in an expected way, but failing to adjust for \hat{p}_j in the linear model has little effect (and, if any, seems to improve the approach). These differences between the linear and logistic models could well be due to differences in the behaviour of linear and non-linear models; nevertheless, these findings make it difficult to conclude with certainty that \hat{p}_j definitely acts as a confounding variable in these circumstances.

Table 24 below provides results for a repeat of these analyses on the 1958BC dataset with SNP spacing of 20. As can be seen, the type I error rates for all approaches are now raised above the expected level and, hence, even though the logistic regression approach with an offset and the GEE independence model are both immune to heteroscedasticity, some characteristic of these data remains problematic. This finding supports the conclusions from Section 2.7.2 that LD is indeed a problem in these data. For instance, in Table 23 – where LD is unlikely to be a problem because the dataset consists of SNPs that are widely spaced across the genome, the logistic regression approach with an offset and the GEE independence model both yield approximately correct type I error rates. However, here – where the SNPs are more closely spaced – both of these approaches now yield an elevated type I error rate.

Analysis	Group	Mean (\hat{b})	Mean [$\text{Var}(\hat{b})$]	Reject. H_0 (5% level of sig.)	Discriminatory Analysis (% correct)
Linear Regression	Case	1	0.0494	0.9809	100
	Control	-1	0.0494	0.9788	100
	Neither	-0.0015	0.0494	0.1129	-
Logistic Regression w/out Offset	Case	3.9792	1.2870	0.8978	100
	Control	-4.0430	1.2872	0.9021	100
	Neither	-0.0395	1.2796	0.1083	-
Logistic Regression w/ Offset	Case	5.1578	1.6560	0.9694	100
	Control	-5.1572	1.6556	0.9675	100
	Neither	-0.0079	1.6449	0.0763	-
GEE Independence Model	Case	1	0.0612	0.9701	100
	Control	-1	0.0612	0.9679	100
	Neither	-0.0015	0.0622	0.0764	-

Table 24: Results for analyses of the 1958BC using SNP spacing of 20 and sampling only individuals with southern UK ancestry. For individuals in the case or control groups the % rejections of H_0 represent power for each approach; for individuals in “neither” group the % rejections of H_0 represents the type I error rate for each approach.

A further aspect of the results that implies that LD is present in this dataset concerns the estimates of the regression coefficient. Estimates of the regression coefficient are consistent here with those presented in Table 23, so the elevated type I error rates must be due to a problem with the variance. LD is likely to lead to underestimation of the variance of the regression coefficient

and, thus, these results (i.e. in Table 24) are consistent with the pattern of results that would be expected in the presence of LD.

Given that LD is suspected in these data, the problems seen here should not be surprising, since all the approaches described so far assume independent observations. The following section attempts to overcome the problems posed by LD by investigating whether an alternative specification of the correlation structure in a GEE model allows a sufficient correction for the correlation between SNPs. A further GEE model is thus proposed, using a *first-order autoregressive* (AR-1) correlation structure.

2.7.4 Adjusting for LD

As seen in the previous section, where GEE models employ the *robust* variance estimate, they successfully adjust for the effects of heteroscedasticity. However, GEEs can also be used to provide consistent estimates of regression coefficients and their standard errors in correlated data. Although GEEs do not generally require the correlation matrix to be accurately specified, the GEE independence model fitted in the previous section assumes that all observations are completely independent. As Table 24 shows, this seems to be too strong an assumption to make in correlated data, because the model yields an elevated type I error rate when the data are suspected to be in LD. In this section I attempt to better approximate the true correlation structure by specifying a first-order autoregressive (AR-1) correlation structure in a GEE model.

2.7.4.1 GEE Model with AR-1 Correlation Structure

In the context of the SNP data used in these analyses, the AR-1 correlation structure specifies that SNPs that are in close proximity to one another within a *cluster* (see next paragraph) are more highly correlated with each other than more distantly located SNPs. Within each cluster, the correlation between the j^{th} and the k^{th} SNP is λ_{jk} , and the AR-1 correlation structure specifies that $\lambda_{jk} = \lambda^{|j-k|}$. Hence, neighbouring observations, i.e. that have distance, $|j-k| = 1$, have correlation λ , and the correlation between any other pairs of observations within the same cluster is λ to the power of $|j-k|$. As with any GEE model, no between-cluster correlation is assumed, so any pairs of observations in different clusters are assumed to be independent.

For the independence estimating equation approach described in Section 2.7.3.2, all the observations are ordered into individual clusters and there is no need to estimate any correlation parameters because the correlation matrix is an identity matrix. However, alternative correlation structures require the structure of the observations to be specified and they require a correlation parameter (or a set of correlation parameters) – such as λ – to be estimated. The structure of the observations is specified by defining clusters. Observations should be clustered in some orderly and logical way. For example, a clinical trial taking multiple measurements on each patient would probably order repeat observations on the same individual together in a cluster – possibly in time order. As the observations we consider here all relate to a single individual in any given test (i.e. of that individual's presence within a mixture or study), there is no such obvious way in which to cluster the observations. GEEs usually

require at least 50 clusters to reliably estimate the correlation parameters (and, hence, to provide good estimates of the parameter variances), however, and it is thus necessary to find some way to split the observations into clusters (Paik, 1988; Yan *et al.*, 2004). As the correlation between SNPs tends to occur only within chromosomes, clustering the observations by chromosome (and by position on the chromosome) is the most logical approach. However, as there are only 22 chromosomes (ignoring the sex chromosomes X and Y), this would provide too few clusters in which to adequately fit the GEE model. The simulations that follow therefore do not adopt this approach. Rather than forming 22 large clusters of observations – one for each chromosome – I instead create many smaller clusters of SNPs. Each cluster consists of SNPs ordered by base position, so SNPs closer together within a cluster should be more highly correlated than SNPs further apart (if correlated at all). Due to computational limitations – where larger size clusters require greater computation time – I initially create clusters of only 20 observations (greater cluster sizes are, however, tested in Section 2.7.4.3). An implication of clustering the SNPs in this way is that some SNPs on the same chromosome – which may be in LD with each other – will be separated into different clusters; the assumption of between-cluster independence may thus be violated here. Despite this constraint, I assess how well the approach performs in Section 2.7.4.2, and I examine the influence of cluster size in Section 2.7.4.3.

2.7.4.2 Testing the GEE AR-1 Model

The 1958BC data with SNP spacing of 20 is initially used to test the GEE model with an AR-1 correlation structure. Given that the GEE independence model

and the logistic regression model with an offset both yield elevated type I error rates in these data, low to moderate levels of LD are suspected to be present in this dataset. As with earlier scenarios, only individuals from southern UK regions are used in these analyses (i.e. to avoid violating the assumption of co-ancestry), and only the first 4,577 SNPs from the dataset are used (i.e. so the results should be more comparable to those reported in sections 2.6 and 2.7).

The simulation method applied here is the same as that used in Section 2.7.3.3. Hence, in each simulation run, 100 southern UK individuals from the 1958BC dataset are randomly sampled without replacement into both arms of a hypothetical case-control GWAS, and 100 individuals are randomly sampled into a test group of individuals under the null hypothesis. The terms q_j and \hat{p}_j are derived for each SNP as described previously (see Section 2.5.1). In each simulation run, every individual within each of the three test groups is tested in turn for presence in the study. One hundred simulation runs are performed in total.

The GEE AR-1 model is applied here using cluster sizes of 20. The first 20 spaced SNPs form the first cluster; the next 20 SNPs form the second cluster; and so on, up until all of the SNPs to be used in the test are included within clusters. Up to 229 clusters are therefore formed for each individual (although the precise number of clusters varies between individuals depending on the amount of missing data). Such clustering allows for LD over a range of 400 SNPs in the full, un-spaced dataset.

Table 25 below shows results for the GEE model with an AR-1 correlation structure. Since the AR-1 model can conveniently be compared to the other

approaches, results for the linear regression, logistic regression with \hat{p}_j included as an offset, and GEE Independence model approaches are also included, with respect to power and type I error. For the AR-1 approach, although estimates of the regression coefficient are very similar (but not identical) to those of the linear regression and GEE independence approaches, the variance is, on average, substantially greater. For instance, compared to the linear regression approach, the variance of the AR-1 approach is around 40% greater, and compared to the GEE independence model the variance of the AR-1 approach is around 20% greater. Consequently, the GEE AR-1 model yields a marginally lower power than the GEE independence and the linear regression models (i.e. power ≈ 0.96 Vs 0.97 and 0.98 respectively) but, more importantly, the GEE AR-1 model also yields the correct level of type I error here (i.e. 5%). As the GEE AR-1 model is the only approach so far that corrects the type I error rate in correlated data this is a distinct advantage over the other approaches. Moreover, even the (small) loss of statistical power it suffers over the other approaches can potentially be recuperated because this approach is able to utilise both independent SNPs and SNPs that are correlated (unlike the other approaches, which only perform adequately if the SNPs used are independent). Hence, many more SNPs are potentially available for use to this method.

Analysis	Group	Mean (\hat{b})	Mean [$\text{Var}(\hat{b})$]	Reject. H_0 (5% level of sig.)	Discriminatory Analysis (% correct)
Linear Regression	Case	1	0.0493	0.9835	100
	Control	-0.9999	0.0493	0.9817	100
	Neither	-0.0092	0.0493	0.1106	-
Logistic Regression w/ Offset	Case	5.1532	1.6450	0.9716	100
	Control	-5.1549	1.6460	0.9692	100
	Neither	-0.0476	1.6344	0.0735	-
GEE Independence	Case	1	0.0609	0.9725	100
	Control	-0.9999	0.0609	0.9706	100
	Neither	-0.0092	0.0620	0.0732	-
GEE AR-1	Case	1	0.0725	0.9585	100
	Control	-0.9998	0.0723	0.9562	100
	Neither	-0.0093	0.0738	0.0507	-

Table 25: Analyses of the 1958 Birth Cohort data with SNP spacing of 20. Results are based on 100 simulation runs. 100 individuals with southern UK ancestry are sampled without replacement into the case and control groups of a simulated GWAS, and 100 individuals are sampled into a group of individuals who are not in the study. Linear Regression, Logistic Regression with an offset term, and the GEE Independence model are all performed in the usual way. The GEE AR-1 model is performed using a cluster size of 20.

To test how the GEE AR-1 approach performs both in weakly correlated and independent data the above analysis has been repeated using the 1958BC datasets with SNP spacing of 33 and 100, as well as a simulated dataset of truly independent SNPs based on the simulation method described in Section

2.6.1 (results not shown). In each analysis the GEE AR-1 approach performs similarly to the above analysis of the 1958BC dataset with SNP spacing of 20. The correct type I error rates are obtained for all datasets, with only modest reductions in power compared to the other approaches. GEE models with an AR-1 correlation structure thus perform well in this context regardless of whether or not LD is suspected, or whether there is heteroscedasticity. Furthermore, in these scenarios the approach performs well despite its assumption of between-cluster independence. In the analyses of the 1958BC datasets with SNP spacing of 20 and 33 there is likely to be correlation between SNPs within different clusters, but the correct type I error rates are still obtained.

These results are consistent with those reported by Clayton (Clayton, 2010), who also found that LD could be adjusted for without having to be precise about the specific nature of the correlation. However, where Clayton proposes estimating the correlation structure using an additional sample of individuals, the GEE AR-1 approach does not require any additional data and is computationally far simpler, as no method for estimating large, sparse matrices is required. Furthermore, although Clayton presents graphics showing that the variance of his Bayes factor can be controlled by adjusting for LD, the Bayes factor itself tends to be biased downward of its true value.

2.7.4.3 Influence of Cluster Size

In the previous section the GEE model with an AR-1 correlation structure performs well both in datasets with independent SNPs and in datasets with suspected low levels of LD. Although the cluster size is fixed at 20 in the reported results, additional tests using increased cluster sizes of 50

observations were also performed (results not shown) and produced the same pattern of results. The GEE approach, thus, seems relatively robust to the cluster size and to the precise nature of the specified correlation structure. Given that GEE approaches assume no between-cluster correlation it is perhaps surprising that the method performs so well in these data. For instance, there *is* likely to be LD between SNPs in different clusters in these datasets, but GEE models can only account for the correlation *within* clusters. The specified correlation structure can therefore only be expected to account for a proportion of the overall correlation between SNPs and, hence, especially when the cluster size is small, the specified correlation structure is likely to be a poor approximation of the actual correlation between SNPs. The datasets considered so far have consisted of relatively widely spaced SNPs, however, and any violations in the assumption of between-cluster independence are likely to have been minor. This section aims to explore how well the GEE AR-1 approach performs in a more extreme dataset where strong LD between SNPs is likely. The 1958BC data is again used for these analyses, but instead of spacing SNPs across the genome, every typed SNP on a particular stretch of a chromosome is selected for use (see next paragraph). The focus of these analyses is to determine how well the GEE AR-1 model performs when there is realistically strong LD and, in addition, whether performance can be improved (if necessary) using larger cluster sizes. The AR-1 correlation structure is likely to be a poor representation of the true correlation structure in data with strong LD, but larger cluster sizes *should* better approximate the correlation, i.e. because each cluster accounts for the correlation between greater numbers of SNPs.

As with the simulations in sections 2.7.3.3 and 2.7.4.2, only individuals with southern UK ancestry are selected for these analyses. Similarly, the analysis is limited to 5,000 SNPs (compared with around 4,600 SNPs previously). In contrast to earlier scenarios, however, all of the SNPs selected here are from one chromosome: the analyses use the first 5,000 SNPs on chromosome 14 and, hence, will be affected by strong LD. As usual, in each simulation run 100 individuals are randomly sampled without replacement into the case and the control arms of hypothetical case-control GWAS and, in addition, 200 of the remaining individuals are randomly sampled to be test individuals in neither group. In each simulation run, every individual within each of the test groups is tested in turn for presence in the case-control GWAS using the original linear regression approach, the GEE independence model, and the GEE AR-1 model. The GEE AR-1 model is performed using cluster sizes of 20, 50, 100 and 200. Because the computation time is increased substantially when performing the analyses with larger cluster sizes, all simulations conducted in this section are based on only 10 simulation runs. Monte Carlo estimates of the regression coefficient and its variance, and of the type I error and power, thus, should only be considered approximate in this section. Results are summarised in Table 26 below.

Approach	Cluster Size	Group	Mean (\hat{b})	Mean [Var(\hat{b})]	Reject. H_0 (5% level of sig.)
Linear Regression	NA	Case	1	0.0453	0.881
		Control	-0.9936	0.0453	0.878
		Neither	0.0055	0.0453	0.4030
GEE Independence Model	NA	Case	1	0.0611	0.854
		Control	-0.9936	0.0611	0.848
		Neither	0.0055	0.0627	0.3195
GEE AR-1	20	Case	1.0012	0.1464	0.687
		Control	-0.9895	0.1473	0.689
		Neither	0.0094	0.1500	0.1265
GEE AR-1	50	Case	0.9988	0.1841	0.633
		Control	-0.9927	0.1842	0.618
		Neither	-0.0103	0.1890	0.0890
GEE AR-1	100	Case	0.9955	0.2001	0.589
		Control	-0.9921	0.1978	0.586
		Neither	0.0261	0.2039	0.0805
GEE AR-1	200	Case	0.9964	0.2016	0.593
		Control	-0.9994	0.2012	0.606
		Neither	0.0007	0.2041	0.0765

Table 26: Results for analyses of 1958BC data – chromosome 14. Rejections of H_0 represents power for case and control individuals, but represents type I error for “neither” individuals.

For the two approaches that assume independence, i.e. the linear regression and GEE independence approaches, the type I error rates here are vastly inflated above the nominal level (e.g. type I error = 0.4 and 0.32 respectively at the 5% level of significance). Although these approaches yield the expected

mean estimates of the regression coefficient, they seemingly underestimate the variance of the regression coefficient. For instance, in comparison to the GEE AR-1 approach, estimates of the variance are around 50% to 80% smaller for the approaches that assume independent observations (e.g. variance ≈ 0.045 - 0.06 for the linear regression and GEE independence approaches respectively Vs 0.15 - 0.2 for the GEE AR-1 approach). These results are to be expected in correlated data, where a failure to account for the correlation between SNPs will lead to biased estimates of the variance.

For the GEE AR-1 approach the type I error rates are also elevated above the nominal level but, as predicted, there appears a clear relationship between the cluster size and the performance of the method. For example, the type I error rate decreases from approx. 0.125 when the cluster size is 20 , to approx. 0.075 when the cluster size is 200 . Although estimates of the regression coefficient are very similar irrespective of the cluster size, the variance of the regression coefficient increases here with the cluster size and, consequently, the type I error rate decreases towards the nominal level with larger clusters. This supports the prediction that larger cluster sizes provide better approximations of the true correlation structure in these models. Nevertheless, as seen in Section 2.7.4.2, when the data are only weakly correlated, the GEE AR-1 approach performs well even with small clusters. Furthermore, even though none of the GEE AR-1 analyses performed here yield the correct type I error rates, all perform considerably better than the approaches that assume independent observations.

In addition to the results shown in Table 26, I also tested the GEE AR-1 model with even larger cluster sizes to investigate whether the correct type I error rates could eventually be attained in these data. However, increasing the cluster sizes further greatly increased the computation time (e.g. a cluster size of 500 took at least 150 seconds to run per individual test on my desktop PC), and simulations testing numerous individuals tended to crash. Thus, I managed to test only small numbers of simulations with these larger cluster sizes. Provisional results for a cluster size of 500 based on only two simulation runs (testing 200 individuals under the alternative hypothesis and 200 individuals under the null hypothesis in each run), shows no further improvement in lowering the type I error rate compared with the best analysis in Table 26 (i.e. which uses a cluster size of 200). However, the analysis using a cluster size of 500 forms only 10 clusters in total because the analysis is limited to using only 5,000 SNPs. It has been suggested that GEE approaches require at least 50 clusters to adequately estimate correlation structures (Yan *et al.*, 2004) and, hence, this analysis uses an insufficient number of clusters. This may be the reason why this analysis performs no better than the analysis with a cluster size of 200, but it has not been possible to verify this explanation (e.g. by using a greater number of SNPs) due to the computational limitations. Further investigation into the influence of cluster size would be useful, but this may have to be conducted in future when computational power has improved and when more efficient algorithms may be developed. It is worth noting, nonetheless, that a real application of this method requires only a single or small number of individuals to be tested. Hence, should further simulations definitively demonstrate the benefit of using large cluster sizes for these models, it may be

possible to use larger numbers of SNPs and larger cluster sizes, i.e. in the knowledge that the test need only be performed a few times.

A key feature of the findings reported in this section concerns the assumption that the GEE AR-1 approach corrects analyses for the effects of LD by imposing an approximate correlation structure on the models. In fact, in a further analysis, I performed a GEE model with an independence correlation structure, but where the observations were clustered in the same way as reported for the GEE AR-1 approach (i.e. in clusters of 20 observations). The results for this analysis were somewhat surprising, because very similar results were obtained compared to the GEE AR-1 model with the same clustering. This analysis, thus, implies that it is the clustering of observations that is actually key to allowing for correlation in these models, rather than the nature of the specified correlation structure itself.

This finding actually supports the description of GEEs provided in sections 2.7.3.2 and 2.7.4.1, in that GEEs *should* be robust to inaccurate specifications of the correlation structure. Furthermore, although the results for this additional analysis markedly improve on the results from the GEE Independence model fitted prior to here (i.e. which uses clusters of single observations), the GEE AR-1 approach remains marginally better in terms of efficiency (i.e. power). The full implications of this finding, i.e. that the clustering of observations is so important, go beyond the scope of this project, but there may be interest in exploring this issue further in later work.

2.8. Ancestry

We have so far seen that the Visscher *et al.* linear regression approach is dependent on compliance to the key assumptions of homoscedasticity (Section 2.7.3) and independent observations (Section 2.7.4). However, the straightforward extension of the method using a generalised estimating equation (GEE) with a first-order autoregressive correlation structure (AR-1) can handle violations in either of these conditions without any adverse effects. In principle, these results thus imply that not only are SNP allele frequencies sufficiently informative to identify participants from genome-wide association studies (GWAS) but, furthermore, a reliable method to identify individuals using SNP data is tractable. Until now, however, we have not considered the implications of the assumption of co-ancestry on these methods.

Co-ancestry is an important prerequisite of these methods because it ensures that the two groups compared in a test are truly comparable and, hence, that a test is only influenced by an individual's presence in or absence from a study. Although case-control GWAS are usually well matched in terms of ancestry, subtle differences between the two arms of a study would be difficult to detect (and, thus, difficult to avoid). Visscher *et al.* (Visscher *et al.*, 2009) acknowledge that the linear regression approach is sensitive to differences in ancestry, and they state a threshold – in terms of Wright's F_{ST} statistic (Wright, 1968) – at which population divergence is problematic. Although they state that an F_{ST} value approaching $1/2N_{mix}$ will cause problems, they do not elaborate on this statement in any way. Different methods of deriving F_{ST} exist (Cavalli-Sforza *et al.*, 1994; Balding, 2003), and inconsistent values of F_{ST} between

different real populations are reported in the literature (Cavalli-Sforza *et al.*, 1994; Pritchard *et al.*, 2001; Cardon *et al.*, 2003; Marchini *et al.*, 2004; Heath *et al.*, 2008). As such, the true implications of differences in ancestry remain unclear.

This section examines how sensitive these approaches are to realistic differences in ancestry, with the aim of determining whether even subtle differences in ancestry throw the methods, i.e. in terms of increased error rates.

2.8.1 Previous Findings and Preliminary Results

In the context of the Homer method (Homer *et al.*, 2008), previous work indicates that differences in ancestry can have a major impact upon the type I error rate. For example, in a simulation study (Sampson *et al.*, 2009), the type I error rate approaches 100% when the two test groups differ in ancestry even when only a small percentage of SNPs (e.g. 1% or more) are “ancestry-associated”. Similarly, in the same paper an analysis using data from the International HapMap Project (2003) yields markedly increased type I error rates when a reference group includes even a single individual with different ancestry to the others. These findings therefore suggest that even subtle differences in ancestry are sufficient to hinder methods to test for presence in GWAS using SNP allele frequencies.

The above findings seem to cast serious doubt on the usefulness of these methods in practice, given that perfect co-ancestry is almost impossible to guarantee in real studies. However, given that this may not be such an issue for case-control GWAS, which are often well matched in terms of ancestry,

these findings may not always be relevant. Moreover, the above analyses seem particularly extreme. For instance, the analyses of data from HapMap create reference groups consisting of Utah residents with ancestry from northern and western Europe (CEU individuals) and at least one Tokyo resident with Japanese ancestry (JPT individuals). Real case-control GWAS would be unlikely to include individuals from such distinct ancestries because of the well known analytical effects of population admixture (or *population stratification*), i.e. population stratification can lead to both false-positive and false-negative results (Marchini *et al.*, 2004). In the context of case-control GWAS, this is thus an extreme illustration of the effects of population admixture on the Homer method.

The applicability of the simulation study described above can also be questioned. As the authors – Sampson and Zhao (Sampson *et al.*, 2009) – admit, the results from these simulations are likely to be extreme because the allele frequencies for the ancestry-associated SNPs are generated completely independently of one another between the two ancestries (i.e. for corresponding SNPs). In reality, allele frequencies for SNPs that differ by ancestry – and, in particular, between similar ancestries – are likely to be correlated (Reiner *et al.*, 2005). Hence, the findings from this study are likely to exaggerate the real effects of differing ancestry in case-control GWAS.

In an attempt to investigate the validity of the findings from Sampson and Zhao (Sampson *et al.*, 2009), I also conducted a set of simulation studies to investigate the effects of violating the co-ancestry assumption in more moderate (and, hence, realistic) scenarios. I considered several different approaches to

simulating ancestry, and initially aimed to use real data to inform the simulation method. For example, I simulated population allele frequencies for two different ancestries from the observed allele frequencies of different, real ancestries in the HapMap database (2003). These analyses yielded extreme results. High type I error rates were obtained, which seemed to indicate a strong sensitivity to ancestry even when the populations compared were similar. Upon reflection of the results, however, I came to realise that these analyses were invalid. Because these analyses simulated population allele frequencies from real, observed allele frequencies, they, in effect, implicitly assumed that the observed allele frequencies were population frequencies, i.e. they assumed that the observed allele frequencies represented the true, underlying allele frequencies for the populations under study. In truth, the observed allele frequencies are really *estimates* of the population allele frequencies and, hence, will be subject to random variation around the true population values. As such, simulating two populations based on observed allele frequencies – as done in these simulations – leads to exaggerated differences between the two populations. The following section therefore further investigates the effects of violations in the co-ancestry assumption, by undertaking a subsequent simulation study using an alternative simulation approach.

2.8.2 Simulation Study

This simulation study simulates ancestries that differ to various extents, and measures the differences between ancestries using a formulation of Wright's F_{ST} reported in the literature (see Equation 20 below).

Allele frequencies for the first population, p_{1j} , are simulated in the usual way, i.e. by generating allele frequencies from a uniform distribution bounded by 0.05 and 0.95. The p_{1j} are then converted to the log-odds scale, where $\pi_{1j} = \log[p_{1j}/(1-p_{1j})]$. Subsequently, a value representing a hypothetical number of binomial trials is selected. This value, which I shall refer to as “N.scale”, controls the degree of divergence between the two simulated populations, so, for instance, lower values of *N.scale* produce greater divergence between the simulated populations, and greater values of *N.scale* produce less divergence. In these simulations, *N.scale* takes a fixed value in each scenario, and is varied between 100 and 2000 in different scenarios.

An expected number of “successes” for each SNP, r_j , is generated by multiplying p_{1j} by *N.scale*. This allows the standard error of the log-odds of p_{1j} to be derived, where $SE(\pi_{1j}) = \sqrt{[1/r_j + 1/(n.scale - r_j)]}$. Finally, the log-odds of the allele frequency in the second hypothetical population, π_{2j} , is generated by randomly sampling from a normal distribution with mean equal to π_{1j} and standard deviation equal to $SE(\pi_{1j})$. The raw allele frequency for population 2, p_{2j} , is then derived using the inverse-logistic (or *expit*) transformation, i.e.

$$p_{2j} = \frac{\exp(\pi_{2j})}{1 + \exp(\pi_{2j})}.$$

Once the underlying allele frequencies have been derived for the two populations, 100 individuals are simulated into each arm of a hypothetical case-control study, using the allele frequencies for one population to simulate the cases and the allele frequencies for the other population to simulate the

controls. As usual, 100 individuals from each population are also simulated to test under the null hypothesis (i.e. as individuals absent from the study).

As stated above, these simulations measure the divergence between simulated ancestries using F_{ST} . As presented by Cavalli-Sforza (Cavalli-Sforza *et al.*, 1994), the formulation of F_{ST} I use for the j^{th} SNP here is:

Equation 20

$$F_{ST} = \frac{\text{Var}(p_j)}{\bar{p}_j(1 - \bar{p}_j)}$$

In the context of the data we have here, $\bar{p}_j = (p_{1j} + p_{2j})/2$ and $\text{Var}(p_j) = (p_{1j} - \bar{p}_j)^2 + (p_{2j} - \bar{p}_j)^2$. An overall measure of F_{ST} can then be derived by taking the mean F_{ST} value over all s SNPs.

Although Equation 20 uses the underlying allele frequencies, p_{1j} and p_{2j} , any derivation of F_{ST} , in practice, requires use of the estimated (or *observed*) allele frequencies, q_j and r_j (denoting the allele frequency in the mixture and reference group respectively). Measures of F_{ST} will thus be sensitive to the sample sizes of the two groups used to estimate the allele frequencies unless the underlying allele frequencies are known.

As already mentioned, Visscher *et al.* (Visscher *et al.*, 2009) state that population divergence becomes a problem as F_{ST} approaches $1/2N_{\text{mix}}$. This formula ignores the influence of the sample size of the reference group, however, and is therefore ambiguous. Furthermore, it is unclear whether this formula relates to a measure of F_{ST} based on underlying allele frequencies or observed allele frequencies. Visscher *et al.* report a series of simulation studies

in their paper, and, thus, would have had access to the simulated underlying allele frequencies (even though, in practice, these would not be available). Further clarification of the relationship between F_{ST} and the performance of these tests is therefore required. The following simulations investigate this relationship, and derive F_{ST} in each scenario based on both the underlying and the observed allele frequencies.

Table 27 below presents the results for various scenarios based on 5,000 SNPs and 100 simulation runs. As the case and control groups consist of 100 individuals in the reported simulations, the critical value of F_{ST} stated by Visscher *et al.* is $1/(2*100) = 0.005$ here.

N.scale	F _{ST} based on:		Approach	Power	Type I Error
	Underlying AFs	Observed AFs			
2000	0.00025	0.0053	Linear Regression	0.999	0.064
			GEE Independence	0.998	0.057
			GEE AR-1	0.998	0.058
1000	0.0005	0.0055	Linear Regression	0.999	0.075
			GEE Independence	0.999	0.073
			GEE AR-1	0.999	0.074
500	0.001	0.006	Linear Regression	0.999	0.122
			GEE Independence	0.999	0.138
			GEE AR-1	0.999	0.139
300	0.0017	0.0067	Linear Regression	0.999	0.219
			GEE Independence	0.999	0.277
			GEE AR-1	0.999	0.277
100	0.0049	0.0099	Linear Regression	1	0.769
			GEE Independence	1	0.919
			GEE AR-1	1	0.918

Table 27: Results for the simulation study investigating the effects of population divergence. The results show the relationship between the degree of population divergence – as measured by F_{ST} – and the error rates for the Linear Regression, GEE Independence (with cluster sizes of 1), and GEE AR-1 models. Power and type I error rates are based on the 5% level of significance.

The results demonstrate that the measures of F_{ST} based on the observed allele frequencies are consistently higher than those based on the underlying allele frequencies. This is to be expected, as the observed allele frequencies are affected by sampling variation and, hence, should be more divergent than the underlying allele frequencies.

Table 27 also demonstrates that the relationship between F_{ST} and the error rate for the linear regression approach reported by Visscher *et al.* must assume use of the *observed* allele frequencies in each group. For instance, in the scenario where *N.scale* is 2000, F_{ST} based on the observed allele frequencies is approximately equal to the critical value (i.e. $F_{ST,observed} = 0.0053$) and the type I error rates approach the nominal level. On the other hand, F_{ST} based on the underlying allele frequencies is well below the critical value in this scenario ($F_{ST,underlying} = 0.00025$). Thus, one would expect perfect type I error rates here if the formula reported by Visscher *et al.* were to relate to the underlying allele frequencies.

Similarly, in the scenario where *N.scale* is 100, which relates to the most divergent groups simulated, F_{ST} based on the underlying allele frequencies approaches the critical value (i.e. $F_{ST,underlying} = 0.0049$) but the type I error rates are massively inflated (e.g. type I error for GEE AR-1 is 0.918). In contrast, the F_{ST} value based on the observed allele frequencies is approx. 0.01 in this scenario (i.e. twice that of the critical value), so, assuming that the Visscher *et al.* formula applies to the observed frequencies, elevated type I error rates should be expected here.

Further scenarios simulated with different sample sizes for the case-control studies also support the above findings (results not shown), i.e. that the formula provided by Visscher *et al.* relates to a measure of F_{ST} derived using the observed allele frequencies. Nevertheless, all of the simulations reported so far use equal sample sizes for the case and control groups. When unequal numbers of cases and controls were simulated, the formula reported by

Visscher *et al.* was seen to be inaccurate. For instance, in a scenario with 100 cases, 500 controls, and an F_{ST} value of 0.04, type I errors of around 25% were observed. Hence, these findings suggest that the formula reported by Visscher *et al.* to relate the performance of the methods to F_{ST} only applies to equally sized reference and test samples.

Focussing on the more general characteristics of the results shown in Table 27, it appears that the two GEE approaches perform better than Linear Regression when the population divergence is small (e.g. when *N.scale* is 2000) but worse when the populations are more divergent. Moreover, the results highlight the importance of co-ancestry upon the approaches. For instance, the type I error rates are substantially elevated in all scenarios reported in Table 27 other than when *N.scale* is at least 2000. Hence, even fairly minor differences in ancestry may throw the methods.

It is difficult to place the above results in the context of the divergence between real populations because F_{ST} values reported in the literature vary wildly. For instance, the degree of divergence between European populations has been reported as varying between 0.01 to 0.05 (Cavalli-Sforza *et al.*, 1994) and between 0.001 and 0.005 (Heath *et al.*, 2008). The reason for these inconsistencies is likely to be due to differences in the sample sizes used to estimate F_{ST} , and, as such, the more recent study reported by Heath *et al.* (Heath *et al.*, 2008) is likely to be more accurate. Even with these figures, however, it remains difficult to ascertain the true implications upon these methods – in the context of case-control GWAS data – because GWAS are usually well matched in terms of ancestry. In the following sections I therefore

utilise real GWAS data to investigate how the approaches perform in the presence of realistic differences in ancestry. Section 2.8.3 constructs test groups from different regions of the UK (using the 1958BC data) and Section 2.8.4 constructs groups using data from different GWAS cohorts.

2.8.3 Comparing UK Regions

In sections 2.7.2-2.7.4, real data from the 1958BC have been used to form hypothetical case-control GWAS consisting of individuals only from southern UK regions. Here, the 1958BC data is again used, but the two test groups are formed with individuals from two *different* UK regions. In this section, hence, I explore whether possible differences in ancestry between individuals from different regions of the UK are capable of hindering the methods.

As in Section 2.7.2, individuals from each of the twelve sub-regions of the UK recorded in the 1958BC data are combined into one of three larger regions: South UK (consisting of London, Southeast, Southwest and South England), Central UK (consisting of East England, North Midlands, Midlands, and Wales) or North UK (consisting of Northwest England, North England, East & West Ridings of Yorkshire, and Scotland). Three comparisons of regions are therefore made: South Vs Central UK; Central Vs North UK; and South Vs North UK. In each simulation run, a hypothetical case-control GWAS is formed by randomly sampling 100 individuals from one region into one arm of the study and 100 individuals from another region into the other arm; 100 individuals from the same two regions are then also sampled as individuals to test under the null hypothesis. As usual, each individual from each of the groups is tested in turn for presence in the study, and 100 simulation runs are performed for each

comparison of regions. This analysis is based on the 1958BC dataset with SNP spacing of 20. Because of the known issues with heteroscedasticity and LD in these data, only results for the GEE AR-1 approach (using a cluster size of 20) are shown.

Table 28 below clearly shows that any differences in ancestry between individuals from different regions of the UK do not affect the GEE AR-1 model. The powers obtained are similar to those reported in Table 25, i.e. in which all individuals are sampled from South UK, and the levels of type I error are approximately correct. Although the results for the other approaches (i.e. linear regression, logistic regression, and GEE independence) are not provided, the same pattern of results also applies; no further increases in the type I error rate are demonstrated here compared to the corresponding rates shown in Table 25.

Regions	Group	Mean (\hat{b})	Mean [Var(\hat{b})]	Reject. H_0 (5% level of sig.)
South Vs Central	In Study – South	1.0000	0.0667	0.9691
	In Study – Central	-1.0002	0.0668	0.9664
	Not in Study – South	0.0052	0.0698	0.0518
	Not in Study – Central	-0.0078	0.0698	0.0557
Central Vs North	In Study – Central	1.0001	0.0660	0.9735
	In Study – North	-0.9999	0.0661	0.9691
	Not in Study – Central	0.0129	0.0671	0.0516
	Not in Study – North	-0.0040	0.0671	0.0559
South Vs North	In Study – South	1.0002	0.0665	0.9742
	In Study – North	-1.0004	0.0663	0.9718
	Not in Study – South	0.0144	0.0675	0.0495
	Not in Study – North	-0.0152	0.0676	0.0564

Table 28: Comparison of regions in the 1958 Birth Cohort with SNP spacing of 20. In each simulation run, 100 individuals from each region are randomly sampled into one arm of a hypothetical case-control GWAS, and another 100 individuals from each region are test individuals under the null hypothesis. The proportion of rejections of H_0 represents power for the individuals in the simulated case-control GWAS, and type I error for individuals not in the study.

To complement the results in Table 28, I derived sets of F_{ST} values comparing samples of individuals from each region of the UK in the 1958BC. Recall that Visscher *et al.* state that population divergence becomes problematic as F_{ST} approaches $1/2N_{mix}$ (Visscher *et al.*, 2009). Generally, the F_{ST} values derived comparing the different regions approximately matched the critical value deduced from this formula. Furthermore, deriving F_{ST} for individuals from the southern UK regions only also yielded similar values. These results, thus, demonstrate that the 1958BC contains only individuals who are well matched in

terms of ancestry. Furthermore, they show that the population divergence between participants from different UK regions is no greater than that between individuals from the same region. The 1958BC is known to be a representative control group for UK individuals, however. In order to further generalise the results, the following section compares individuals from different UK GWAS cohorts.

2.8.4 Comparing Different UK Cohorts

In this section, individuals from one of three real UK cohorts are sampled into different arms of hypothetical case-control GWAS. This section thus investigates whether the findings from Section 2.8.3, i.e. that any differences between individuals of UK ancestry are insufficient to throw the methods, are applicable across studies. Different studies have different recruitment procedures and could be subject to different biases; thus, even if the participants within each study appear well matched in ancestry, the subtly different characteristics of each study might perturb the methods.

Permission was granted from the WTCCC to access the genotype data from the UK National Blood Service Controls (NBS) and the Coronary Artery Disease (CAD) cases. As described in Section 2.7.1, the 1958BC consists of 1504 unrelated participants from the UK. Similarly, the NBS consists of 1500 unrelated participants from the UK, and the CAD consists of 1988 unrelated, coronary artery disease patients also from the UK. All genotypes used in these analyses are typed on the Affymetrix 500K chip and called in Chiamo-Oxford format. Every 20th SNP on chromosomes 12 to 19 is selected for these analyses (giving 6097 SNPs in total), before 365 SNPs are omitted following

advice in the exclusion files that accompany the data. Also following the advice in the exclusion files, 24 participants are excluded from the 1958BC dataset, 42 individuals are excluded from the NBS dataset, and 62 individuals are omitted from the CAD dataset.

In each simulation run, 100 individuals from a particular cohort are randomly sampled without replacement into one arm of a hypothetical case-control GWAS, and 100 individuals from another cohort are sampled into the other arm of the hypothetical study. As usual, 100 individuals from the two cohorts are also sampled to test under the null hypothesis, i.e. to test individuals who are not in the hypothetical studies. Three scenarios are tested in total comparing individuals from the 1958BC with a NBS group; the 1958BC with a CAD group; and the NBS with a CAD group. One hundred simulation runs are performed in total, with each individual in each group tested in turn for presence in the study in each run. Consistent with previous scenarios, the analyses are limited to the first 5000 SNPs in the datasets. As in Section 2.8.3, results for the GEE AR-1 approach are presented only, because of the known heteroscedasticity and LD in the data.

Table 29 shows that the GEE AR-1 model performs consistently well in these data regardless of the cohorts compared. The results shown here – and, in particular, the type I error rates – are comparable to those obtained in previous scenarios (e.g. in Section 2.7.4) and, hence, any differences in ancestry between participants in these cohorts do not affect the method. Although results for the other approaches are not shown, they perform similarly. For instance, although they yield elevated type I error rates, these are no greater

than in previous scenarios (e.g. in sections 2.7.2 and 2.7.3) and, as previously explained, are likely to be due to heteroscedasticity and/or LD.

Cohorts	Group	Mean (\hat{b})	Mean [Var(\hat{b})]	Reject. H_0 (5% level of sig.)
1958BC Vs NBS	In Study – 1958BC	1.0000	0.0665	0.9677
	In Study – NBS	-1.0006	0.0664	0.9694
	Not in Study – 1958BC	0.0095	0.0676	0.0539
	Not in Study – NBS	-0.0111	0.0676	0.0573
1958BC Vs CAD	In Study – 1958BC	1.0001	0.0645	0.9733
	In Study – CAD	-1.0006	0.0640	0.9728
	Not in Study – 1958BC	0.0276	0.0658	0.0549
	Not in Study - CAD	-0.0320	0.0655	0.0571
NBS Vs CAD	In Study – NBS	1.0001	0.0643	0.9745
	In Study – CAD	-1.0014	0.0649	0.9717
	Not in Study – NBS	0.0284	0.0658	0.0553
	Not in Study - CAD	-0.0299	0.0665	0.0526

Table 29: Comparison of Different UK Cohorts. In each simulation run, 100 individuals from each study are randomly sampled into one arm of a hypothetical case-control GWAS, and another 100 individuals from each study are test individuals under the null hypothesis. The number of rejections of H_0 represents power for the individuals in the simulated case-control GWAS, and type I error for individuals not in the study.

As with the results reported in Section 2.8.4, I derived F_{ST} values for samples of individuals in the above cohorts in order to deduce the degrees of population divergence. Again, regardless of the pair of cohorts being compared, the F_{ST} values tended to approximately equal the critical value deduced from the formula reported by Visscher *et al.* (Visscher *et al.*, 2009). As Table 29 shows no major increases in type I error, these results, hence, support the formula provided by Visscher *et al.*. Furthermore, the results suggest that no major differences in ancestry exist between individuals in these UK cohorts.

2.9. What can be published?

We have seen that under certain conditions it *is* possible to reliably infer presence within genomic mixtures such as GWAS. The key issue underpinning this work is therefore to clarify what information can and cannot be published safely from GWAS. There are two obvious approaches to avoiding identification in the release of summary data from GWAS: (1) release only a limited amount of data, which is insufficient to allow identification; or (2) release data only in forms that are non-identifiable. These approaches are now discussed.

2.9.1 Limiting the number of SNPs

The SecureGenome software developed by Sankararaman *et al.* (Sankararaman *et al.*, 2009) aims to quantify the amount of information that can be released safely from GWAS. Rather than using the Homer method, SecureGenome employs a likelihood ratio test statistic because this – they claim – provides an upper bound on the maximum power achievable by any test. SecureGenome requires the full genotypes to be input for individuals within a mixture (e.g. a GWAS cohort) and within a reference group. It also requires a measure of rank to be input for each SNP (such as a p-value). Hence, it can only be used by study researchers who have access to these data.

By implementing a procedure that omits SNPs that are adjudged to be in LD, a specific set of SNPs – chosen by rank, is output for which allele frequencies can be published at specifiable levels of type I error and power. Using the SecureGenome software, it can thus be argued that the risk of participant

identification can already be sufficiently avoided by publishing only the set of SNPs deemed “safe”. This approach is effectively an intermediate strategy between releasing full summary information (e.g. on a genome-wide basis) and releasing nothing at all. However, the development of an alternative strategy to avoiding participant identification (e.g. which involves releasing only “non-identifiable” data) may allow a full set of summary information to be released.

2.9.2 Other types of summary information

As an alternative to the release of allele frequencies from GWAS, other obvious forms of summary data that could be of use in genomics research include odds ratios, test statistics, and p-values (either exact p-values e.g. $p=0.03$; or binned p-values e.g. $p<0.05$). In its present form, the Homer method cannot be applied using any of these summary measures; however, alternative tests based on a similar principle to the Homer method are likely to be tractable.

One such test has already been proposed by Clayton (Clayton, 2010). This test utilises z-score statistics to compare an individual of interest to the two groups of a case-control study and to a third reference group. These z-scores are, in effect, a measure of association between each SNP and the phenotype of interest (i.e. the disease that defines the case-group). For each SNP the z-score is derived by subtracting the allele frequency in the case group, \bar{x}_2 , from the allele frequency in the control group, \bar{x}_1 , and then dividing this by the square root of $1/N_1 + 1/N_2$ (where N_1 and N_2 are the sample sizes of the control group and the case group respectively). The *sign* of z therefore conveys which group has the higher allele frequency (i.e. if \bar{x}_1 is greater than \bar{x}_2 , z will be positive; if \bar{x}_2 is greater than \bar{x}_1 , z will be negative), and the *magnitude* of z gives an

indication of the size of the difference between the allele frequencies in the two groups. The Clayton test statistic is based on the following notation. For a particular SNP, x denotes the genotype for an individual of interest ($x = 0, 0.5$ or 1); μ denotes the minor allele frequency (MAF) in the reference group ($0 \leq \mu \leq 0.5$); and z denotes the z -score as described above. The test statistic, T , is thus derived as $T = (x - \mu) z$. Under the null hypothesis, T is expected to be zero because there is no correlation between $(x - \mu)$ and z . If the individual is closer to the control group, T will be positive, and if the individual is closer to the case group, T will be negative. This can be explained as follows.

Because μ is the MAF (which is always between 0 and 0.5), $(x - \mu)$ will be positive for individuals with one or two copies of the minor allele, and negative for individuals with no copies of the minor allele. The test is therefore driven by two processes: (1) the sign of $(x - \mu)$ relative to the sign of z ; and (2) the magnitude of $(x - \mu)$ relative to the magnitude of z . The sign of $(x - \mu)$ with respect to the sign of z conveys which of the two groups the individual of interest is closer to. For instance, an individual with no copies of the minor allele will be closer to the group with the lesser MAF, and an individual with one or two copies of the minor allele will be closer to the group with the greater MAF. Hence, if the MAF is greater in the control group (i.e. z is positive), T will be positive for an individual with one or two copies of the minor allele, and T will be negative for an individual with no copies of the minor allele. If the MAF is greater in the case group (i.e. z is negative), T will be negative for an individual with one or two copies of the minor allele, and positive for an individual with no copies of the minor allele. The magnitudes of z and $(x - \mu)$ convey the strength of the information obtained for a given SNP. For instance, intuitively it is clear

that if there is little difference between the MAFs in the case group and the control group (i.e. z is small), a SNP will provide little information to the test. Conversely, if the difference in MAFs in the case and control groups is large, an individual could be significantly closer to one of the two groups than the other (depending on his/her genotype).

Clayton suggests calculating two Bayes factors based on the above statistic to test for presence in a case-control study. Although these Bayes factors are not shown, this work implies that test statistics – and, hence, p-values with appropriate directionality – are also identifiable. It is also remarked, however, that the use of the chi-square statistic, z^2 , cannot be used in place of the z -statistic because z^2 is unsigned (i.e. z^2 does not convey which group has the greater MAF). The sign is crucial in the above tests because it is this that correlates with the individual of interest's genotype when he/she is in one of two groups. Without a sign, no correlation between an individual and his/her group would be possible, and it thus seems logical that summary measures which do not convey the sign should not be identifiable. Appropriate summary measures – which are “un-signed” and which could therefore potentially be released – are to be discussed shortly. Before this, I first aim to prove that merely the sign on its own is also sufficient to identify participants from DNA mixtures.

2.9.3 Sign Test

The simplest form of information that can be used to compare the allele frequencies in a case-control GWAS is the *sign* (i.e. of the difference in allele frequencies between the two groups). This sign conveys which group has the greater allele frequency, but it is completely uninformative of the magnitude of

any difference in the allele frequencies. To denote the sign for the j^{th} SNP, a variable, $Sign_j$, is created ($Sign_j = 1$ if the MAF is greater in the cases; $Sign_j = 0$ if the MAF is greater in the controls). Note that SNPs for which the MAFs in the two groups are equal are omitted from the analysis. SNPs are also omitted if the allele with the lesser frequency (i.e. the minor allele) differs between case and control groups.

A test based on the sign of the difference in allele frequencies between the case and control groups of a GWAS involves creating a binary variable, D_j , to denote which group the individual of interest is “closer” to for each SNP ($D_j = 1$ if the individual is closer to the case group; $D_j = 0$ if the individual is closer to the control group). As has been described in the previous section, if the MAF in the control group is greater than the MAF in the cases (i.e. $Sign_j = 0$), an individual with one or two copies of the minor allele will be closer to the controls than the cases (hence $D_j = 0$), and an individual with no copies of the minor allele will be closer to the cases than the controls ($D_j = 1$). In the opposite situation where the MAF is greater in the cases, an individual with one or two copies of the minor allele will be closer to the cases ($D_j = 1$), and an individual with no copies of the minor allele will be closer to the controls ($D_j = 0$). The distance of the individual of interest from each group is therefore dichotomised by the variable D_j .

For a set of independent SNPs, and for two groups of similar size and similar ancestry, it seems reasonable to assume that, under the null hypothesis, mean D_j will be approximately 0.5. This is because if an individual is in neither group, he/she would be expected to be “closer” to each group approximately the same

number of times. If an individual is in the case group, mean D_j should be significantly greater than 0.5 because he/she will be closer to the case group more often than the control group. If an individual is in the control group, mean D_j should be significantly less than 0.5 because he/she will be closer to the control group more often than the cases. A two-tailed binomial test can thus be used to test the null hypothesis that $D_j = 0.5$. Where W is the sum of all D_j , and M is the total number of SNPs for which the MAFs in the case and control groups are different, the test statistic, T , is therefore:

$$T = W \sim \text{Bin}(M, p=0.5).$$

Some simple simulations were conducted to evaluate the power and type I error rate for this sign-test. Based on the simulation method in Section 2.4, 20,000 SNPs were generated with population allele frequencies randomly sampled from a uniform (0.05, 0.5) distribution. Case and control groups consisting of 500 individuals were then simulated, with a further 500 individuals simulated in neither group. Each individual from the case group was tested for presence in the cases, and each individual from the control group was tested for presence in the controls. The individuals in neither group were each tested for presence in both the case and control groups.

At the 5% level of significance, approximately the correct level of type I error was yielded (i.e. approx. 2.5% in each tail), with a power of approximately 80%. In comparison to the ROC curves presented in Section 2.4.2, it thus appears that this sign-test is only marginally less powerful than the original Homer method. Although this test is also subject to the same set of assumptions and

constraints as the Homer method, it confirms that identification is possible when only the signs of differences in allele frequencies between two groups are published. Hence, any statistic that conveys these signs – such as odds ratios, and even binned p-values – will also be identifying if the direction of effect is known.

2.9.4 Implications

As we have seen, these tests are driven by the correlation between the signs of differences in allele frequencies between two groups, and the genotypes of an individual of interest. Any aggregated statistics that convey these signs are therefore potentially identifying. By omitting details regarding the directionality of an effect, however, identification via these means can be prevented. For example, odds ratios that do not state the allele that is associated with an outcome would be safe. Similarly, two-sided p-values that hide this information would also be safe. These “un-signed” statistics could still be informative to the genomics research community, and a shift back in practice towards publishing aggregate data would represent an improvement on the current demands for researcher disclosure. Caution must remain in the publishing of these data, however. Systematic coding of aggregate statistics – such as by always presenting associations between the minor allele and the disease group of a study, could allow the sign to be accurately guessed. Furthermore, publishing additional information such as standard errors may allow raw allele frequency data to be derived by solving sets of simultaneous equations. An additional concern is whether resources such as the HapMap Project could be used to infer signs or directions of effect. It is not obvious how these resources could

be used to systematically infer, for example, whether a particular allele frequency is higher in a case group or a control group. Any estimates based on current HapMap allele frequencies would also be relatively imprecise (i.e. because they would be based on no more than 90 subjects for each ethnicity). As increasingly advanced resources – such as HapMap 3 and the 1000 Genomes project - become available over the coming years, more work will be needed to reassess the threat to participant confidentiality. In light of these future developments, increasingly complex ways in which to breach participant anonymity may become possible.

2.10. Discussion

The findings in this chapter generally concur with results reported by others (Homer *et al.*, 2008; Braun *et al.*, 2009; Jacobs *et al.*, 2009; Sampson *et al.*, 2009; Sankararaman *et al.*, 2009; Visscher *et al.*, 2009; Clayton, 2010). The Homer *et al.* test (Homer *et al.*, 2008), although dubious from a methodological point of view, raises important concerns regarding the privacy of data in genetic epidemiological studies. Under certain conditions, SNP allele frequencies *are* informative of an individual's presence in or absence from a study and, hence, at least to some extent, the reactions of the NIH and the Wellcome Trust to remove these data from the Web (see Appendix B.1) seem justified.

As long as the key assumptions of co-ancestry and independent observations are upheld, it does seem possible to identify a participant in a GWAS cohort. Typically, several thousand SNPs are required to identify an individual in a study of one or two hundred participants. Larger studies consisting of a few

hundred individuals generally require an order of ten thousand SNPs. Even greater sized studies – for example, consisting of five hundred to a thousand subjects – would require several tens of thousands of SNPs. It is important to note that the simulation results provided throughout this chapter are generally conservative for the sample sizes investigated, because relatively small numbers of SNPs have been simulated. Practical use of these methods might allow use of many more SNPs and, hence, greater powers would potentially be achievable. Note, however, that increased sample sizes would require greater power.

We have seen that use of these methods in practice requires a reference group as well as a test mixture (such as a GWAS cohort) and a genomic profile for an individual of interest. The implication of this for the Homer test is that the specification of the null distribution is not always accurate, particularly when the sample sizes of the two groups differ. In contrast, the framework introduced by Visscher *et al.* (Visscher *et al.*, 2009) avoids problems regarding the sample sizes of the two groups compared in a test. Visscher *et al.* propose a linear regression approach that – from a statistical point of view – is more coherent than the Homer method. Furthermore, it also generally out-performs the Homer test. As Section 2.6 shows, however, this linear regression approach consistently yields marginally elevated type I error rates.

2.10.1 Heteroscedasticity

The small elevations in type I error rates that are consistently observed for the linear regression approach are caused by heteroscedasticity in the model's error terms. This, in turn, is due to modelling genotype data, which is inherently

binomial in nature. The linear regression approach, thus, has an incorrect variance function for these data. Section 2.7 outlines straightforward extensions to the linear regression approach to allow for heteroscedasticity. The proposed logistic regression approach addresses the problem by correctly modelling the variance function (see Section 2.7.3.1), whereas the proposed GEE approach addresses the problem by using a *robust* estimate of the variance (see Section 2.7.3.2). Both of these approaches seem to adequately deal with heteroscedasticity, and yield the correct type I error rates when no other model assumptions are breached.

2.10.2 The implications of linkage disequilibrium (LD)

A further characteristic of real datasets that impacts upon the approaches is LD. Section 2.7 demonstrates that LD is a problem in real data even when analyses use only every 20th or every 33rd SNP. Only the dataset with SNP spacing of 100 seems relatively free from LD and, hence, these findings indicate that LD can range over up to 100 SNPs in an Affymetrix 500K scan. Any approach that assumes independent observations provides biased estimates of the variance in datasets containing SNPs in LD, and, consequently, yields elevated type I error rates.

The GEE AR-1 approach introduced in Section 2.7.4 seems effective at allowing for low to moderate levels of LD. This approach clusters neighbouring observations together and allows for correlation between observations within the same cluster. In the datasets with SNP spacing of 20 and 33, the GEE AR-1 approach corrects the type I error rate and yields only slight reductions in power over the other approaches. When the SNPs are more densely located,

however, even this approach is affected. Although it performs better than the other approaches in a dataset containing un-spaced (and, thus, highly correlated) SNPs, its type I error rate remains above the expected level. This is likely to be due to the somewhat arbitrary clustering of the observations here, and the fact that GEE models assume no between-cluster correlation. For example, if the cluster size is 100, although correlation between SNPs within each cluster is accounted for, any *between* cluster correlation, say, between SNPs 90 to 110 would only be partially accounted for, i.e. the correlation between SNPs 90 to 100 would be accounted for in one cluster, the correlation between SNPs 101 to 110 would be accounted for in another cluster, but any correlation between SNPs in the different clusters would not be taken into account.

An extension of the work reported in this chapter could investigate better strategies for clustering the observations for the situation we consider here. The obvious approach would be to cluster the observations by chromosome and, hence, avoid between-cluster correlation by clustering any observations in LD with one another in the same clusters. As has already been mentioned, however, GEE models typically require at least 50 clusters of observations (Yan *et al.*, 2004) but humans have only 22 homologous chromosome pairs. An alternative approach, therefore, could be to select SNPs from two or three distantly located regions in each chromosome, and to cluster the SNPs from each region together. This approach would be likely to minimise any between-cluster correlation, as all observations in different clusters would be located either on different chromosomes or in different regions of a chromosome. Alternatively, as has been used in sections 2.7 and 2.8, thinning the data by

only selecting SNPs spaced across the genome provides a means for avoiding LD partially or even altogether.

As touched upon at the end of Section 2.7.4, the results from these analyses emphasise the importance of the clustering structure in a GEE model. In the presence of correlation, the *clustering* structure seems to be far more important than the specification of the *correlation* structure. In essence, this observation concurs with Clayton (Clayton, 2010), who also finds that correlation can be adjusted for without having to be precise about the exact nature of the correlation structure. In contrast to the Clayton method, however, the GEE approach avoids having to estimate a large, sparse covariance matrix, and, hence, is computationally simpler.

2.10.3 Ancestry

The results in Section 2.8 show that even minor violations in the co-ancestry assumption can have a major impact upon the type I error rates of the approaches. Nevertheless, the analyses based on real GWAS data suggest that case-control studies, which typically carefully match participants for ancestry, may often be sufficiently similar for the approaches to be unaffected.

Visscher *et al.* (Visscher *et al.*, 2009) provide a formula stating that the linear regression approach is not useful if Wright's F_{ST} statistic (Wright, 1968) approaches $1/2N_{mix}$. The simulation results in Section 2.8 broadly support this formula; any scenarios in which F_{ST} (estimated from the observed allele frequencies in each group) exceeds this critical threshold do have elevated type I errors, and the scenarios in which F_{ST} approximately equals the threshold are

untroubled. F_{ST} therefore seems to provide a reasonable means for assessing whether two equally sized groups are sufficiently similar to one another to allow a test for presence within the studies to be performed. However, the present formula is inaccurate when the groups are of unequal size and, moreover, no obvious extension of the formula is available for this situation.

2.10.4 Forensic use of the tests

Although the primary focus of this work has been on the implications of these methods on the publication of data from GWAS, the original Homer *et al.* paper (Homer *et al.*, 2008) was motivated by a forensic application. A number of similarities between the two situations exist and, in some respects, they may be considered analogous. However, one key discrepancy between the two situations is that, in GWAS, each individual contributes equally to the allele frequencies, whereas in the forensic mixture, the percentage of DNA contributed by each individual may vary. As Egeland *et al.* (Egeland *et al.*, 2010) show, when different individuals contribute unequally to a mixture, the ability of the test to reliably infer presence within the mixture is undermined. As such, the utility of these methods in the forensic application is ultimately compromised.

2.10.5 Conclusions

We have seen that a GEE extension to the Visscher *et al.* linear regression approach (Visscher *et al.*, 2009) appears the most suitable method for inferring the presence of an individual in GWAS using SNP allele frequencies. However, simple rules advising on precisely what can and cannot be published safely

from GWAS remain difficult to surmise. For instance, the ability to reliably identify an individual in a study depends on a number of factors. Clearly, the sample size of the test group is crucial; however, the sample size of the reference group is important too. The extent of any correlation between the SNPs used in a test also affects the method – both in terms of the power and the type I error rate. Furthermore, the degree of population divergence between the two test groups ultimately determines whether the method is even useful at all. The information that can be made available is, thus, highly study dependent.

As Section 2.9 shows, other types of information – such as odds ratios or p-values – also appear identifying if they convey the sign of any effect. Conversely, although “unsigned” p-values appear safe, further analysis is required to establish whether there is a risk of being able to infer the signs (e.g. by using existing resources on the Web or in the literature). Further dialogue is also required to determine whether this information would even be useful at all.

Most of the results currently being published from GWAS stem from large-scale meta-analyses of – typically – tens of thousands of subjects. Hence, publication of the results from these studies, at least for the time being, appears safe. The number of SNPs required to identify an individual in studies of these sizes would be likely to run into the hundreds of thousands or even millions, and strong LD between the SNPs would therefore be inevitable. At present, no approach has been shown to be capable of adequately dealing with major correlation between SNPs. With future advances in computing, however, methods to allow for strong LD may be tractable; indeed, I suggest one such

approach in Section 2.7.4.3. Nevertheless, a further characteristic of these meta-analyses may ultimately hinder any attempts to identify. Meta-analyses typically include studies from a number of different countries, and may therefore include participants of numerous different ancestries. The assumption of co-ancestry, thus, may be impossible to uphold between any two, different meta-analyses. It therefore seems perfectly safe to publish at least the top several thousand SNP hits from these studies.

Chapter 3.

A new approach to data synthesis: DataSHIELD

3.1. Introduction

Recent advances in genetic epidemiology have typically been achieved using huge numbers of participants (e.g. (Levy *et al.*, 2009; Newton-Cheh *et al.*, 2009; Stahl *et al.*, 2010)) and the increasing complexity of genomic research will require even greater sample sizes (Burton *et al.*, 2009). Typically, rather than acquiring the required sample size in a single, large-scale study, consortia have been formed by teams of researchers representing several smaller studies (e.g. Global BPgen (Newton-Cheh *et al.*, 2009) and CHARGE (Levy *et al.*, 2009) consortia). The required sample sizes have therefore been achieved by combining data from multiple resources. Studies participating in these consortia are usually distinct from one another – both from a funding and a researcher perspective – and are often distributed across national borders and boundaries. Data synthesis techniques are therefore necessary to amalgamate the results from the different studies.

Currently, the most common approach to synthesising data across studies in genomics is *study-level meta-analysis* (SLMA). SLMA involves analysing each

study separately before combining the results in a weighted average (see Section 3.1.1.1). Because SLMA requires all analyses to be specified *a priori*, it can be considered to be a restrictive approach to analysis. A more flexible approach is *individual-level meta-analysis* (ILMA), which involves combining the individual (or *patient*) level data from each study to form a single, large dataset (see Section 3.1.1.2). There are often ethical and legal constraints to the sharing of these data, however, and, for this reason, the use of ILMA is sometimes prohibited altogether (see Section 3.1.2).

This chapter proposes a new approach to data synthesis that improves the flexibility of analyses, while potentially circumventing some of the ethical and legal restrictions typically associated with data sharing (Wolfson *et al.*, 2010). Section 3.1.1 describes the existing approaches to data synthesis in more depth, before some of the ethical and legal issues involved in the sharing of data are discussed in Section 3.1.2. Section 3.2 outlines a potential solution to the problem, which we call “DataSHIELD”, and Section 3.3 illustrates two different uses of DataSHIELD. Finally, Section 3.4 presents a discussion of the key issues involved.

3.1.1 Existing approaches to data synthesis

The synthesis of results across different studies is predominantly performed by meta-analysis (Hedges *et al.*, 1985; Sutton *et al.*, 2000), and meta-analysis is broadly undertaken in one of two ways (Sutton *et al.*, 2008). As stated in the previous section, SLMA involves analysing each study separately and pooling the summary statistics (for example, regression coefficients and their standard errors) in a weighted average. In contrast, ILMA involves pooling the *individual-*

level data from each study together to form, in effect, a single large dataset; this is then analysed using conventional statistical methodology (such as by fitting a linear or generalised linear model).

Either approach to meta-analysis requires any participating studies to be sufficiently similar to one another (Fortier *et al.*, 2010). For example, it is usually only meaningful to pool data from studies that have measured both the outcome and any primary exposures of interest in the same (or very similar) ways (Borenstein *et al.*, 2009; Wallace *et al.*, 2009). All participating studies must have a comparable estimate of the exposure effect sizes upon the outcome. Studies that are similar in these respects are often referred to as being “harmonised”, and “harmonisation” is thus an important prerequisite of any meta-analysis (Burton *et al.*, 2010). The two approaches to meta-analysis are now described further.

3.1.1.1 Study Level Meta Analysis

SLMA is often performed retrospectively, where, for example, a comprehensive literature review is initially undertaken to identify any suitable studies, before the required summary statistics are collated from published literature. In the context of genome-wide association studies (GWAS), however, a more prospective approach to SLMA is usually involved. For instance, consortia of genetic epidemiological studies are typically led by one of the participating research groups. This lead group decides upon the analyses to be performed (e.g. by specifying the models to be fitted) and communicates these requirements to each of the other participating groups in an “analysis plan”. Subsequently, each group analyses their own study data according to the

analysis plan, produces the required set of results, and then returns these to the lead group. Once all summary statistics have been returned, the lead group then performs the “overall” analysis by taking a series of weighted averages (see Section 3.2.3).

SLMA typically involves fitting one of two different types of model, and the choice of model to use is usually determined by assessing whether the effect sizes estimated from each study can be assumed to be homogeneous. The effect sizes are described as homogeneous if the differences between them can be attributed solely to sampling error (otherwise known as random variation) (Sutton *et al.*, 2000). Different methods exist to assess whether this assumption is reasonable (e.g. (DerSimonian *et al.*, 1986; Higgins *et al.*, 2003)) and, if so, a fixed effect model may be used. Fixed effect models assume that a single true effect size exists and, hence, that any differences between studies are due to sampling error (Borenstein *et al.*, 2009). If the effect sizes are inconsistent between studies – or, in other words, if there appears to be *between-study heterogeneity* – a random-effects model is usually preferred. Random-effects models assume that the estimated effect sizes represent a random sample of the true effect sizes; in particular, they acknowledge that the true effect may be different in different studies but all effects have a common distribution. In situations where there is between-study heterogeneity, one must remain cautious in the interpretation of the final results even where a random-effects model has been used. This is because between-study heterogeneity can be indicative of differential bias between studies. Alternatively, between-study heterogeneity can also indicate the possible presence of an underlying interaction with an exposure of interest. Any such interaction with an exposure

is potentially important, and typically requires further investigation. For instance, exploratory analyses may need to be performed (such as subgroup analysis (Gelber *et al.*, 1987)), and additional models may need to be fitted to investigate an interaction (Riley, *et al.*, 2010). These analyses generally require deriving further summary statistics from each study or accessing the individual-level data from each study and, hence, can be difficult to perform in an SLMA. In this sense, SLMA is thus an inflexible approach. SLMA requires all analyses to be specified *a priori*, and provides little scope for conducting exploratory or follow-up analyses without repeating the process *de novo*. As we shall see in sections 3.2 and 3.3, DataSHIELD specifically addresses these issues by better enabling the performance of exploratory analyses during data synthesis.

For SLMAs of GWAS, the random effects model is generally recommended for use (Ioannidis *et al.*, 2007; McCarthy *et al.*, 2008). By definition, GWAS involve testing a large number of genetic variants, and it is thus not feasible to assess each variant for homogeneity in its effect size across all participating studies. Random effects models tend to be more conservative than fixed effect models and, thus, can be considered the safer approach to analysis. Similarly, heterogeneity between-studies may actually be expected in an SLMA of GWAS – particularly where studies are included from different countries – because of the possible effects of population stratification. Hence, on this basis too, random effects models do seem the most appropriate choice of approach to synthesise data from different GWAS.

Both fixed effect and random effects models take regression coefficients and their standard errors from each study and combine them in a weighted average.

The regression coefficients are weighted according to precision, so that the studies with greatest precision in their estimate of the effect size have greater weighting in the analysis. The estimated effect size from each study, thus, is weighted by the inverse of its variance. As precision is largely determined by sample size, larger studies tend to have greater weighting in an SLMA.

3.1.1.2 Individual Level Meta Analysis

In contrast to SLMA, an ILMA is performed by combining the individual-level data (sometimes known as individual patient data (Sutton *et al.*, 2008)) from all the participating studies. The resulting single, large dataset can then be analysed using conventional methodology (such as linear or logistic regression) as if it were data from a single study. Crucially, between-study heterogeneity can be accounted for in an ILMA by including study-specific terms in the model (Riley, *et al.*, 2010). These terms can be handled either as fixed or random effects, depending upon the assumptions one wishes to make (Higgins *et al.*, 2001; Whitehead *et al.*, 2001; Riley, *et al.*, 2010).

ILMA requires prior ethical and legal permission to share the raw, individual-level data from any participating study. Studies can usually only share these data if they have the full, informed consent to do so from all study participants. This, realistically, must be gained at the outset of a study and, hence, any studies that do not already have this consent can be restricted in terms of their ability to share. Restrictions over the sharing of data can thus prohibit the performance of an ILMA altogether. As sections 3.2 and 3.3 show, however, use of DataSHIELD potentially offers a means for benefiting from the properties of an ILMA without ever having to share any individual-level data. .

3.1.1.3 Comparison of the approaches

As we have seen, both SLMA and ILMA provide a means of synthesising data from multiple studies. SLMA takes summary statistics from each study and combines them in a weighted average, whereas ILMA pools the individual-level data from each study and analyses the combined dataset. Figure 20 below illustrates the two approaches schematically.

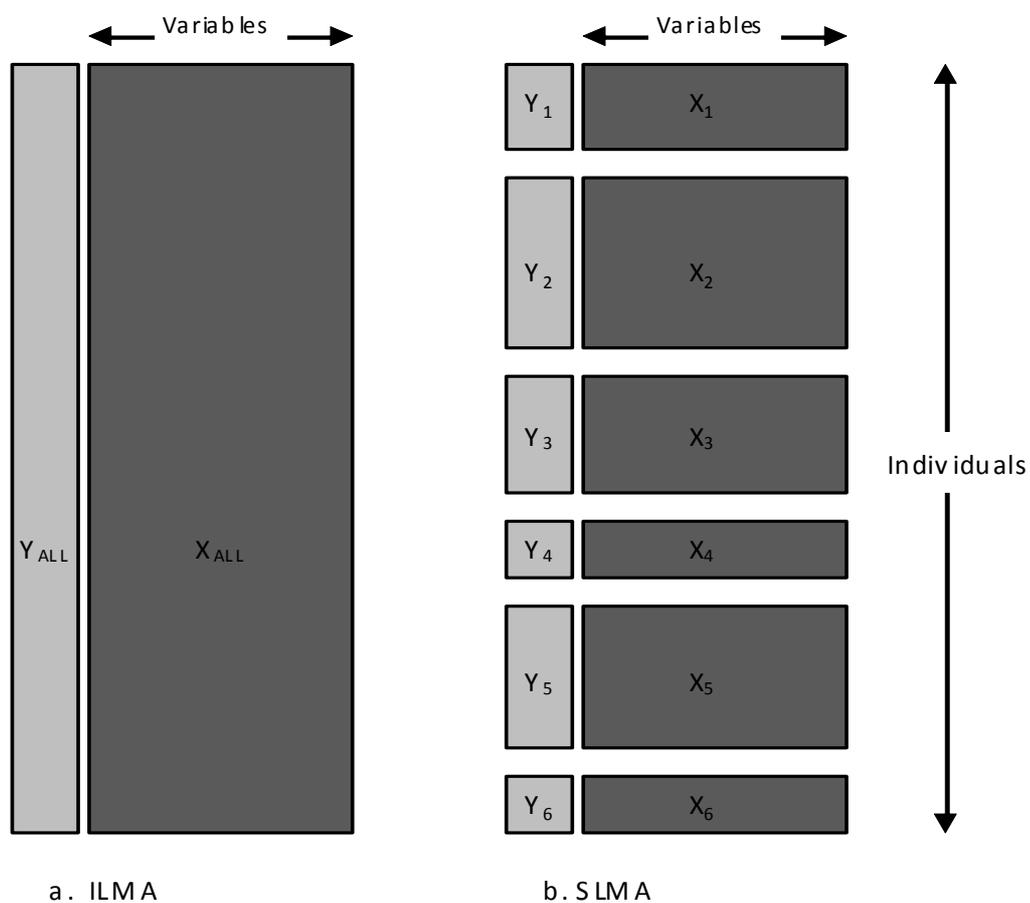


Figure 20: Schematic representation of the ILMA (a) and SLMA (b) approaches. Each diagram displays the outcome vector, Y , in light grey and the covariates – or *design* – matrix, X , in dark grey. For ILMA, the combined dataset is analysed to provide overall estimates of the regression coefficients and their standard errors. For SLMA, each distributed dataset is analysed separately to produce study-specific estimates of the regression coefficients and standard errors. These are then pooled in a weighted average to provide the overall set of results. This diagram is taken from the DataSHIELD paper (Wolfson *et al.*, 2010).

Despite the differences in how they are executed, both approaches often perform very similarly to one another (Olkin *et al.*, 1998; Sutton *et al.*, 2008). For instance, the two approaches applied to fixed effect models of continuous, normally distributed data can potentially yield *identical* results, while models fitted to other data types (e.g. binary data) will produce similar (although not necessarily identical) results.

The key advantage SLMA has over ILMA is that SLMA does not require sharing any of the raw, individual-level data from any participating studies. Thus, as will be discussed further in the following section, there are fewer ethico-legal constraints to the use of SLMA compared to ILMA. On the other hand, as we have already seen, the key advantage ILMA has over SLMA is greater flexibility (Riley, *et al.*, 2010). ILMA has instant access to the individual-level data from all participating studies, so exploratory analyses are quick and easy to perform. In the context of GWAS, specialist procedures such as haplotype analysis and genotype imputation require access to individual-level data and, thus, are also easier to perform with ILMA compared with SLMA (McCarthy *et al.*, 2008). To summarise, ILMA is the ideal approach to data synthesis but, as we shall now see, because it requires access to the individual-level data from any participating study it cannot always be used.

3.1.2 Ethico-legal issues surrounding the sharing of data

Funding bodies actively encourage the sharing of data in genomics (Foster *et al.*, 2007), but research teams face both an ethical and legal obligation to protect the confidentiality of study participants (Kaye *et al.*, 2009; Resnik, 2010). There is, hence, a conflict between the scientific goals of a study and the moral

and legal duty to ensure the protection of participant privacy. Scientifically, researchers have a responsibility to gain as much insight from a given study as possible, which can often only be achieved by sharing and combining data with other research groups (Trinidad *et al.*, 2010). Yet any failure to adequately protect personal data faces potentially damaging ramifications. In the UK, for instance, the Data Protection Act 1998 regulates the “processing” of personal data, and can severely punish any failure to take “reasonable” steps to prevent data loss. Where data are shared, those originally responsible for protecting the data could find it difficult to maintain control of privacy standards. A reluctance to share is therefore not unusual.

In Chapter 2 we saw that participant confidentiality can be difficult to guarantee in GWAS even when aggregate data are only ever released. Clearly, individual-level genomic data is even more sensitive, and could directly allow the identification of study participants. Individual-level data is therefore subject to strict regulation, and its use often restricted to the original researchers who carried out a study (Wallace *et al.*, 2009). Even in situations where the sharing of data, in principle, is allowed, the practicalities involved in actually doing so can be prohibitively long drawn. For example, approval from both a scientific committee and an ethical review panel must be obtained (Malfroy *et al.*, 2004; Eisenstein *et al.*, 2009). Similarly, the sharing of data across borders is often prohibited completely (Kaye, 2005), but, even when, in theory, it is allowed, the different legislation between different countries can severely hamper attempts to share (Zink *et al.*, 2008).

In general, the value of sharing data and resources between different research groups does seem to be recognised by the wider public in addition to scientific communities (Trinidad *et al.*, 2010). However, the existing ethico-legal constraints reflect real dangers and concerns regarding the misuse of data and, as such, do not seem likely to be relaxed. Attempts have therefore been made to address the need to share data in genomics in spite of these ethico-legal stipulations. One idea that has been proposed is to introduce researcher IDs that, once approved, allow researchers access to a pre-specified set of databases (P3G Consortium *et al.*, 2009; Resnik, 2010). Data from any participating study could thus be synthesised by researchers who have permission to access all of the required datasets. The main problem with this approach, however, is that it does not get around the need to obtain prior informed consent from study participants to share the data with researchers who may not have been part of the original research group (Greely, 2007). As noted above, a number of studies do not have this permission and, hence, would not be able to participate in such a scheme. Thus, although it may be possible to gain the required consent in future studies, this idea is of only limited use for existing studies that do not have this consent.

The ethico-legal constraints on the sharing of data have generally prevented ILMA from being used to synthesise results from GWAS and, consequently, SLMA has been the favoured approach. SLMA allows data from each study to be analysed only by the original researchers, and involves passing on only summary statistics from each dataset. SLMA can thus be performed without infringing any of the legal stipulations that restrict data use. As has been discussed in Section 3.1.1, however, SLMA is not an ideal approach to analysis

because it requires all analyses to be pre-specified. Section 3.1.3 describes the characteristics required for an improved approach to synthesising data in genomics.

3.1.3 What is needed?

Any new approach to synthesising data across studies in genomics must address the scientific need for greater flexibility in analyses while taking account of the ethico-legal constraints on the sharing of individual-level data. Ideally, analyses need to be both specified *and* executed from a single hub, to allow a lead group of researchers the flexibility to explore data as they reasonably see fit. However, data privacy must be maintained by sharing only summary data from each study; any new approach must not permit access to individual-level data beyond the original research groups.

The synthesis of results from different studies, as is the focus of this chapter, is analogous to a situation described in the technometrics literature as an analysis of “horizontally-partitioned” data (Karr *et al.*, 2007). As opposed to “vertical-partitioning” of data, where different attributes relating to the same individuals are distributed among different databases, “horizontal-partitioning” is where different databases contain records for different individuals, with the same attributes measured on each individual. Figure 20, shown earlier, illustrates how the “overall” or “pooled” dataset displayed on the left can be considered to be “horizontally-partitioned” when the data from each study resides in separate, distinct locations, as shown on the right. Under some circumstances, analyses of horizontally-partitioned data can be performed to provide identical results to an analysis of the combined dataset without transporting or sharing data

beyond the original data sites (Karr *et al.*, 2007). Indeed, as described in Section 3.1.1.3, an SLMA fitted to continuous, normally distributed data can potentially do this. However, the conventional SLMA lacks flexibility. Producing the same result in an analysis of a horizontally-partitioned dataset to an analysis of the pooled dataset clearly seems desirable but is not always possible, and a more flexible approach than SLMA is required. Subsequent sections introduce a new approach to data synthesis, which meets these requirements. Crucially, this new approach also avoids sharing any individual-level data beyond the original research groups involved in each participating study. The approach – called “DataSHIELD” – therefore addresses each of the needs of a new strategy to data synthesis described here.

3.2. DataSHIELD

Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual levEL Databases – or “DataSHIELD” (Wolfson *et al.*, 2010) – provides a flexible and secure means of synthesising data between studies. As Section 3.2.1 describes further, DataSHIELD involves executing analyses from a single hub – the “Analysis Computer” (AC) – which provides the power and flexibility to perform exploratory analyses as and when needed. The analysis itself, however, is largely performed on separate “Data Computers” (DCs), which reside locally at the base of each study and hold the individual-level data for that study. Each DC analyses the study data held locally and sends only a set of summary statistics to the AC. Thus, no individual-level data is ever shared with anyone other than the authorised researchers in a given study group. Upon receiving summary statistics from all DCs, the AC synthesises the data to

provide an “overall” set of (possibly interim) results. At no point is the AC required to physically access any of the individual-level data from any participating study (other than, perhaps, a study performed locally, which the researchers are already authorised to access). DataSHIELD therefore allows analyses to be centrally controlled while, at least in principle, avoiding the violation of any of the ethico-legal stipulations regarding the sharing of data.

3.2.1 What is DataSHIELD?

DataSHIELD (Wolfson *et al.*, 2010) is a new approach to synthesising data between studies that encompasses both a dedicated IT infrastructure and the use of specialist statistical algorithms. This chapter outlines the core principles underpinning DataSHIELD. Details of the algorithms required to perform two different statistical analyses in DataSHIELD are provided, and the fundamental characteristics of the IT system needed to implement DataSHIELD are described. Work implementing DataSHIELD is ongoing, however, and the development of the software wrapper needed to control and automate the communications between the AC and the DCs, for example, remains in progress. Specific details about the IT system are beyond the scope of this project and are not included in this chapter; nevertheless, Section 3.2.5 describes the key features required in this system, and Section 3.4.2 includes a general discussion of the IT requirements.

The fundamental premise of the DataSHIELD approach can most clearly be demonstrated diagrammatically. Figure 21 below illustrates the required IT infrastructure and the processes involved in a hypothetical example.

Analyses begin from the centre of the network – the AC – which, in the context of a consortium of GWAS, would usually be controlled by the lead research group. The AC specifies the analyses to be performed (for example, by specifying the model to be fitted), and transmits these requirements to each DC, in parallel, in a short block of computer code. Each DC, which contains the individual-level data for a particular study, then runs the code and derives a set of summary statistics for that study, before returning these to the AC. Once summary statistics for each study have been received, the AC amalgamates the results.

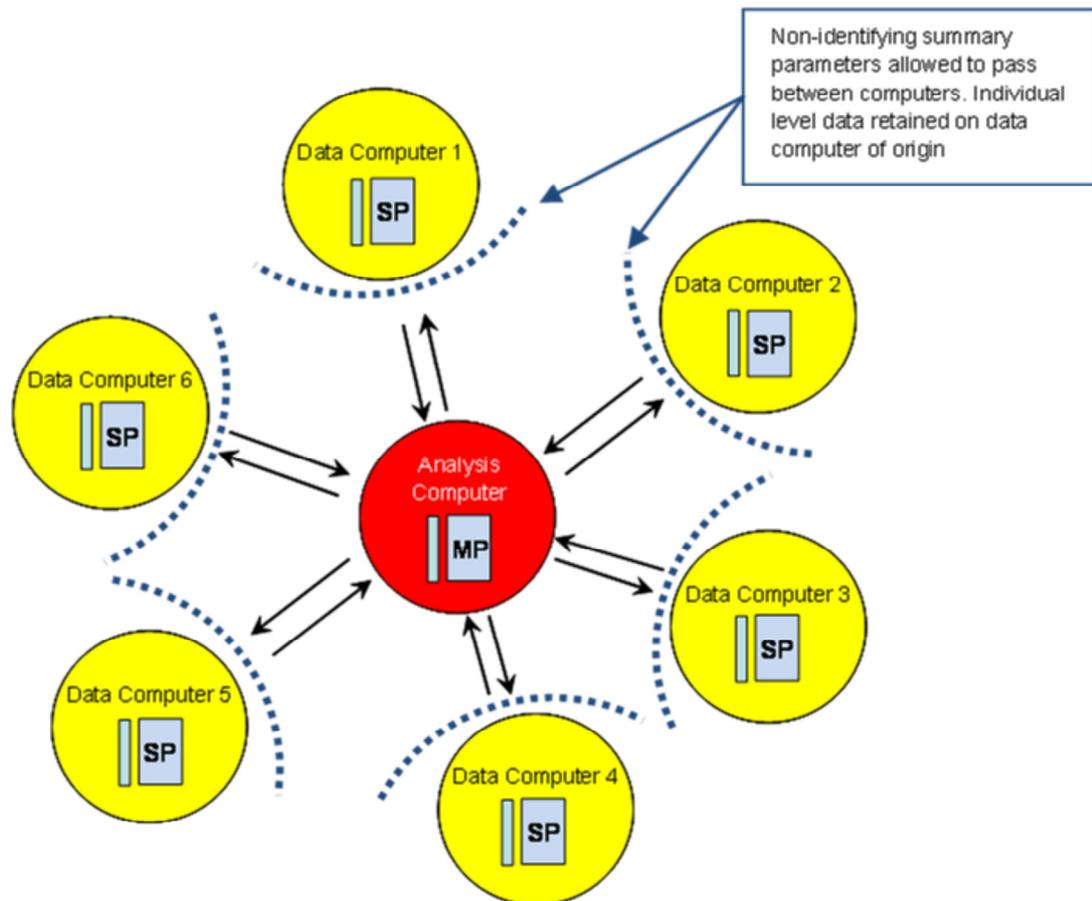


Figure 21: Illustration of the DataSHIELD IT infrastructure (taken from (Wolfson *et al.*, 2010)). The arrows in the above illustration represent the flow of information between the AC and the DCs. Analyses are specified at the Analysis Computer (AC) and transmitted to each Data Computer (DC). Each DC then performs a set of slave processes (SPs), such as fitting a model to the particular dataset held locally, before returning a set of summary components to the AC. Subsequently, the AC performs a set of master processes (MPs), which involve synthesising the summary components from each study. Where applicable, the AC then returns a set of update parameters to each DC to commence a subsequent iteration of the model fitting procedure.

The way in which the AC synthesises the results from the participating studies depends upon the nature of the analysis being undertaken. For example, simple descriptive statistics (such as the average percentage of individuals on antihypertensive treatment in each study) can typically be combined in a weighted average (see Section 3.2.2). Similarly, summary statistics for a linear model are also combined in a weighted average following, in effect, the same

model fitting procedure as in a conventional SLMA (see Section 3.2.3). For generalised linear models (GLMs), the model fitting procedure is more complex, because an iterative algorithm must be used. GLMs are usually fitted using the Iteratively Reweighted Least Squares (IRLS) algorithm (McCullagh *et al.*, 1991), which repeatedly refines estimates of the parameter coefficients and their standard errors until they stabilise. As will be explained further in Section 3.2.4 (and as will be demonstrated in Section 3.3.2), the IRLS algorithm can be fitted in DataSHIELD to guarantee identical results to an analysis of the pooled dataset – or ILMA – if such were possible.

DataSHIELD can be described as a “parallelised” analysis, because it performs analyses on all DCs in parallel. The key advantage of conducting analyses in this way is that it allows exploratory analyses to be fitted virtually instantaneously. For instance, DataSHIELD involves sending only short blocks of computer code from the AC to each DC, and only summary statistics from the DCs to the AC. Hence, utilising the speed of modern internet connections, the process can, in principle, be achieved with little lag compared to a conventional SLMA or ILMA. Various exploratory analyses can therefore be performed from the AC almost as if it has full access to the individual-level data from each study. Note, however, that the performance of some exploratory analyses will be problematic. Some exploratory analyses could allow the retrieval or inference of individual-level data, which could lead to the identification of study participants. For example, most simply, one could request to view a particular row of data from a study, which would provide individual-level information on a particular participant. Alternatively, one could home in on an individual by restricting a summary to specific conditions; for instance, if the date of birth was

known for a known participant, it might be possible to extract the individual level data for this individual either directly (e.g. “print SBP for individuals with DoB = 1/10/1983”) or indirectly (e.g. “print SBP for individuals with DoB \leq 1/10/1983; then “print SBP for individuals with DoB $<$ 1/10/1983”). As such, certain safeguards need to be implemented in DataSHIELD to protect it against possible misuse and, ultimately, to retain the key benefits it potentially offers in terms of data privacy. Some possible safeguards are discussed in Section 3.2.5.

An important prerequisite of DataSHIELD is that all participating studies are harmonised (Fortier *et al.*, 2010; Wolfson *et al.*, 2010), and that all datasets are coded in the same way. For instance, to avoid errors, all variables must be coded with the same names in all datasets. Similarly, the units used to measure any applicable variables must all be the same. Section 3.2.5 discusses other key requirements of DataSHIELD further.

3.2.2 Deriving descriptive statistics in DataSHIELD

Descriptive statistics can easily be obtained using DataSHIELD to summarise and explore the data from participating studies. Usually this will involve a request from the AC to each DC to derive a particular summary statistic of interest for each study. For example, the AC may wish to derive the overall mean number of participants receiving antihypertensive treatment in all the participating studies, or the mean percentage of participants with systolic blood pressure (SBP) greater than 140 mmHg. Once a DC has derived a statistic, it transmits this back to the AC – which, upon receiving a result from all DCs, then derives an “overall” statistic by taking a weighted average. For instance, for the

k^{th} study ($k=1,\dots,s$), the weight, w_k , is equal to the sample size in that study, n_k , divided by the total sample size in all participating studies, N ; and a descriptive statistic of interest (which, for example, could be a count or percentage measure) is p_k . The overall statistic, p , is then derived by multiplying p_k by w_k and summing over all studies:

$$p = \sum_{k=1}^s p_k * w_k$$

Note that a conventional SLMA would typically derive descriptive statistics in the same way, i.e. by using a weighted average. The only novel aspect of this particular procedure therefore relates to the IT infrastructure DataSHIELD uses. This IT system allows a lead research team the power to derive descriptive statistics from participating studies without delay and without having to rely on a researcher from each study manually providing the requested results. It is therefore advantageous over conventional methods, which would experience this delay.

3.2.3 Fitting a linear model in DataSHIELD

As mentioned above, linear models can also be fitted in DataSHIELD by taking weighted averages. In effect, DataSHIELD thus simply coordinates an SLMA from the AC for the fitting of linear models. Nevertheless, the advantage that this provides over a conventional SLMA, as stated before, is that it allows different models to be fitted quickly. DataSHIELD therefore provides a capability for conducting analyses quickly and easily from a single research

base, and it makes complex investigations into, for example, gene-gene and gene-environment interactions, tractable.

The fitting of a linear model in DataSHIELD begins from the AC. The AC specifies the model to be fitted and transmits this to each DC in the form of a short block of computer code. This code contains instructions, in an appropriate programming language (such as “R”), to fit a linear model including the chosen regression terms. Once the DCs receive these instructions they automatically run the code and, hence, fit the appropriate linear model to the study data held locally. Subsequently, each DC transmits a matrix of results back to the AC, which simply contains estimates of the regression coefficients and their standard errors for the corresponding study.

Once the AC has received the results from every DC, the overall analysis is conducted. This firstly involves taking a weighted average of the regression coefficients following the same procedure as outlined in Section 3.2.2. For example, using the same notation as introduced in Section 3.2.2, the overall estimate (or weighted average) of a particular regression coefficient of interest, \hat{b}_k , is derived by multiplying the weight for the k^{th} study, w_k (as shown in the example in Section 3.2.2 above), by the corresponding estimate of the regression coefficient for that study, \hat{b}_k , and summing over all studies:

$$\hat{b} = \sum_{k=1}^S \hat{b}_k * w_k.$$

The AC then pools the standard errors between the participating studies. For a particular regression coefficient, the standard error (SE) for each study, SE_k , is converted to a precision (or inverse variance) by squaring it and taking its

inverse. The precisions are then summed over all studies to derive an “overall” precision, \hat{p} :

$$\hat{p} = \sum_{k=1}^s 1/SE_k^2.$$

Finally, the overall precision is converted to an overall SE, \widehat{SE} , by inverting it and taking its square root:

$$\widehat{SE} = 1/\sqrt{\hat{p}}.$$

Section 3.3.1 contains a simulated data example illustrating the use of DataSHIELD for a linear model as described above, and Appendix C1 provides R code for implementing the model in the example.

3.2.4 Fitting a GLM in DataSHIELD

As with a linear model, generalised linear models (GLMs) can also be fitted in DataSHIELD to produce the same set of results as an ILMA. In order to guarantee this property, DataSHIELD involves a customised model fitting algorithm, which is based on the IRLS algorithm (see Section 3.2.4.1 below) but which allows for the horizontally-partitioned nature of the data. Note, therefore, that a linear regression model, as demonstrated in the previous section, can also be fitted in DataSHIELD using this customised algorithm, by using an appropriate (i.e. identity) link function. The full procedure and algorithm used to fit a GLM in DataSHIELD is outlined in Section 3.2.4.2.

Fitting a GLM involves an iterative procedure that repeatedly refines estimates of the regression coefficients and the standard errors until they stabilise.

DataSHIELD therefore also requires use of an iterative procedure for the fitting of a GLM. As with any DataSHIELD analysis, the procedure begins from the AC by specifying the model to be fitted. As usual, the regression terms to be included in the model must be specified. In addition, GLMs also require the link function of the model (e.g. a logistic link) to be specified, and they require specification of a set of initial values for the regression coefficients. For most regular link functions, these initial coefficient values can simply be set to zero, however, so it seems reasonable to have these values set to zero by default (Venables *et al.*, 2002).

Once the model specifications are transmitted to the DCs, the first iteration of the model fitting algorithm is performed. Each DC is instructed to derive an *expected information* matrix and a *score* vector (see Section 3.2.4.1) using the initial values of the regression coefficients and the study data held locally, before passing the two components back to the AC. The AC then sums the components from each study – deriving an overall expected information matrix and an overall score vector – before it completes the first iteration by incorporating these in the IRLS algorithm equation (see Section 3.2.4.2). This provides estimates both of the regression coefficients and the standard errors. Once these have been obtained, the AC performs a test for convergence (see equations 24 and 25 in Section 3.2.4.1 below). If the algorithm converges, the regression coefficients can be said to have sufficiently stabilised and, hence, the current parameter estimates represent final estimates. If the algorithm fails to converge, the entire procedure must be repeated, but using the current estimates of the regression coefficients in place of the initial values of the regression coefficients. For instance, the AC passes the current coefficient

values to each DC, which derives an updated expected information matrix and an updated score vector using these values. These are then passed back to the AC so the overall analysis can be performed, before the algorithm is again tested for convergence. Typically, this procedure requires around four iterations to achieve convergence.

3.2.4.1 The IRLS Algorithm

The IRLS algorithm is an iterative method of maximum likelihood estimation for GLMs (Aitkin *et al.*, 1989; McCullagh *et al.*, 1991). It relates closely to the Newton-Raphson procedure, but uses the *expected information matrix* to update the regression parameters at each iteration instead of the *observed information matrix* (which the Newton-Raphson procedure uses). The general form of the IRLS algorithm for the r^{th} iteration is:

Equation 21

$$\hat{\mathbf{b}}_{r+1} = \hat{\mathbf{b}}_r + \mathbf{I}(\hat{\mathbf{b}}_r)^{-1} \mathbf{s}(\hat{\mathbf{b}}_r)$$

where $\hat{\mathbf{b}}_r$ is the vector of estimated regression coefficients at the start of the current iteration, $\mathbf{I}(\hat{\mathbf{b}}_r)$ is the estimated *expected information matrix*, $\mathbf{s}(\hat{\mathbf{b}}_r)$ is the *score vector* and $\hat{\mathbf{b}}_{r+1}$ is the updated vector of regression coefficients at the end of iteration r that provides the coefficient values to be used in the *next* iteration ($r+1$). The inverse of the expected information matrix is the *variance-covariance matrix* of parameter estimates.

In the following example, each component of the IRLS algorithm is derived for a logistic regression model, i.e. to model a binary response.

In the first iteration, and for the i^{th} subject, the linear predictor LP_i is derived by multiplying out the model equation:

$$LP_i = X\hat{\mathbf{b}}_1$$

where X is the design matrix, i.e. the matrix consisting of a column of 1s followed by covariate values, and $\hat{\mathbf{b}}_1$ is the vector of initial values for the regression coefficients.

Fitted probabilities \hat{p}_i are then obtained using the inverse logistic transformation, sometimes known as the expit transformation:

$$\hat{p}_i = \exp(LP_i) / [1 + \exp(LP_i)].$$

The expected information matrix $I(\hat{\mathbf{b}}_1)$ is now estimated:

Equation 22

$$I(\hat{\mathbf{b}}_1) = X^T W_1 X,$$

where W_r (here W_1) is a weight matrix (X is again the design matrix), and the superscript 'T' indicates matrix transposition. The weight matrix is a diagonal matrix with diagonal elements w_i , each equal to

$$w_i^{-1} = V_i \{g'(\mu_i)\}^2,$$

where $g'(\mu_i)$ indicates the first derivative of the link function – here the *logistic* function – and V_i is the variance function for the i^{th} subject. For a logistic regression model, $\mu_i = \hat{p}_i$; $V_i = \hat{p}_i(1 - \hat{p}_i)$; and $g'(\mu_i) = 1/[\hat{p}_i(1 - \hat{p}_i)]$. At the r^{th} iteration W_r is derived from the particular parameter values that pertain at that

iteration, and W_1 is therefore based on the parameter values that apply in the first iteration.

Finally, the score function $s(\hat{\mathbf{b}}_1)$ is derived:

Equation 23

$$s(\hat{\mathbf{b}}_1) = X^T W_1 \mathbf{u}_1,$$

where X and W_1 are as before, and \mathbf{u}_1 is a vector of subject-specific terms (u_i), where $u_i = (y_i - \mu_i)g'(\mu_i)$: $y_i = 1$ if subject i is a case; $y_i = 0$ if subject i is a control.

At the end of the first iteration, $\hat{\mathbf{b}}_2$ is derived from $\hat{\mathbf{b}}_1$ using $I(\hat{\mathbf{b}}_1)$ and $s(\hat{\mathbf{b}}_1)$ via Equation 21.

The next iteration, $r = 2$, is then performed taking $\hat{\mathbf{b}}_2$ as the vector of regression coefficients, generating $I(\hat{\mathbf{b}}_2)$ and $s(\hat{\mathbf{b}}_2)$, and using these via Equation 21 to update $\hat{\mathbf{b}}_2$ to obtain $\hat{\mathbf{b}}_3$. This whole process is repeated successively until convergence is achieved.

In R, the `glm()` function has a convergence criterion that is a function of the residual deviance for the current model, D_r , and the residual deviance for the previous model, D_{r-1} . For instance, the default convergence criterion for `glm()` satisfies the following condition:

Equation 24

$$\frac{|D_r - D_{r-1}|}{|D_r| + 0.1} < \varepsilon,$$

where $\varepsilon = 1e-8$.

The residual deviance must be calculated at each iteration:

$$D_r = 2 [\log L_F - \log L_C],$$

where $\log L_F$ is the log-likelihood for the full (or “saturated”) model and $\log L_C$ is the log-likelihood for the current model. For a logistic regression model with binary (1,0) outcomes, as used in the current example, the log-likelihood for the current model is

Equation 25

$$\log L_C = \sum_{i=1}^n [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) + C]$$

where C is a constant, $n_i = 1$, and all other parameters are as before. As the log-likelihood for the full model is always 0 (Quinn *et al.*, 2002), $D_r = -2 * \log L_C$.

3.2.4.2 Applying the IRLS algorithm to horizontally-partitioned data

In order to replicate the results from a GLM analysis of a complete, pooled dataset in a horizontally-partitioned dataset, DataSHIELD applies the IRLS algorithm in steps. Some of these steps are performed from the AC and some from the DCs. As detailed above, the AC begins the analysis by transmitting the model specifications (i.e. the model terms, the model link, and the initial values of the regression coefficients) to each DC. In the first iteration ($r=1$), the DC for the k^{th} study then derives the expected information matrix, $\mathbf{I}_k(\hat{\mathbf{b}}_{r=1})$, the score vector $\mathbf{s}_k(\hat{\mathbf{b}}_{r=1})$, and the log-likelihood for the current model, $\log L_{Cr=1k}$,

using the locally held study data (all as shown in the previous section). Importantly, these are non-identifying, and do not, in themselves, disclose obvious individual-level information. Subsequently, the summary components from each study are returned to the AC to complete the iteration. To do this, the AC first sums the study-specific components across all studies, i.e. the overall information matrix is obtained as $\sum_{k=1}^S \mathbf{I}_k(\hat{\mathbf{b}}_r)$, and overall score function as $\sum_{k=1}^S \mathbf{s}_k(\hat{\mathbf{b}}_r)$. Following Equation 21, the inverse of the overall information matrix is then multiplied with the overall score vector, and the resulting vector added to the initial values of the regression coefficients to yield updated parameter estimates.

In order to test for convergence, each DC must derive and transmit to the AC a study-specific log-likelihood for each iteration. These are obtained locally following Equation 25. Upon receiving the log-likelihood from all studies, the AC sums these, and derives the overall deviance by multiplying the resulting value by -2. This allows the test for convergence shown in Equation 24 to be performed, such that it provides an identical result to an equivalent analysis of a pooled dataset.

As stated above, if the algorithm converges, the current values of the regression coefficients are taken as the final parameter estimates; however, if the algorithm does not converge, the current values are taken as updated coefficient values and are used in place of the $\hat{\mathbf{b}}_1$ values in the next iteration.

Figure 22 overleaf illustrates the procedures involved in fitting a GLM to horizontally-partitioned datasets using DataSHIELD.

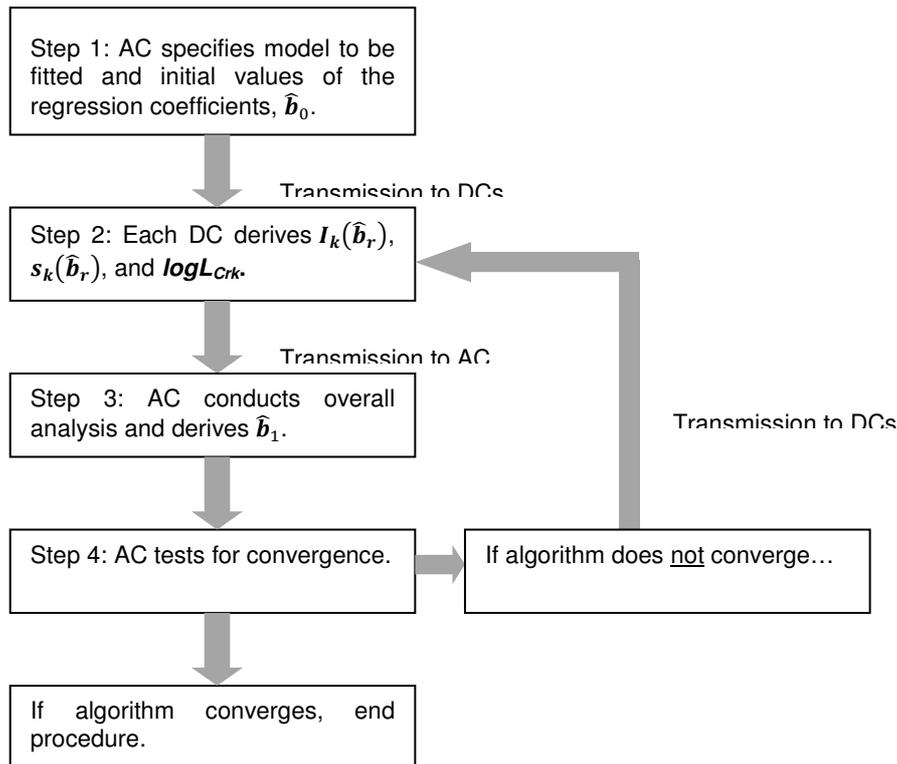


Figure 22: Flowchart representing the processes involved in fitting a GLM in DataSHIELD.

R code for applying a GLM model using DataSHIELD is provided in Appendix C2, in the context of an analysis of a simulated data example described in Section 3.3.2.

3.2.5 Key requirements

This section describes the key requirements of DataSHIELD, largely focussing on the needs of the IT system. Specific details regarding how to implement these requirements are not provided, however. For instance, the low-level code needed to automate the various procedures involved in DataSHIELD will require expertise from specialist computer scientists, and the development of these scripts is therefore beyond the scope of this project. Certain aspects of the implementation of DataSHIELD also require further thought and discussion

before the final IT system can be developed. These features are highlighted in this section, but a more complete discussion of these needs is provided in Section 3.4.

As has already been mentioned, a key requirement of DataSHIELD is that all the participating studies are harmonised, such that the outcome and any explanatory variables of interest are all measured and coded comparably. Therefore, the levels for any categorical variables must be encoded identically; the units for any continuous variables must be the same; and all variables must use the same labels. Such standardisation is, in fact, a fundamental requirement of any meta-analysis, but it could be argued that DataSHIELD requires even more stringent standardisation controls than conventional approaches to data synthesis. This is because DataSHIELD conducts analyses remotely, and prevents the lead researchers from ever accessing any individual-level data themselves. Hence, any mistakes or inconsistencies in the coding of variables could be particularly difficult to identify. Thorough prior consultation with all the participating research groups should therefore be held before undertaking any DataSHIELD analysis, to ensure the correct standardisation of the included datasets.

DataSHIELD requires the creation of a customised IT infrastructure, which must incorporate several key specifications. So as to maximise the flexibility and efficiency of DataSHIELD, once the initial analysis specifications (such as the type of model to be fitted and the terms to include) have been input, all other procedures involved need to be automated. For instance, the DCs need to be programmed to automatically execute instructions received from the AC, and to

pass back any resulting summary statistics. Similarly, the AC then needs to automatically synthesise the results upon receipt of the summary component(s) from each DC. Where applicable, the AC also needs to automatically restart the model fitting procedure to commence any subsequent iteration (this would be required, for instance, in the fitting of a GLM).

Related to the automating of the procedures involved in DataSHIELD is the software wrapper. The precise features to be included in this require some thought, as there could be various strengths and weaknesses associated with the inclusion of different design features. Software will need to be installed both on the AC and on all DCs, perhaps using a different version of the software on each computer type. For the DCs, the software needs to be set up to receive instructions from the AC, and to securely send back results to the AC (see below). The software may also include the code required to perform the various statistical functions of DataSHIELD, so that analysis specifications can be given without the need to transmit the full code for carrying out the analysis with every request.

Essentially, the software for the AC must contain the code to perform the overall analyses for different analysis types, as described in sections 3.2.2 to 3.2.4, for example. Some kind of user interface is also required, so the requirements of an analysis can be specified easily. Furthermore, a facility for outputting the results is required. Results could either be saved, for example, as an object in R, or they could be output as a report.

A key issue that needs to be considered regarding the software for the AC is how much freedom of use to provide to the lead researchers. For instance, if no controls over the derivation of descriptive statistics are put into place, individual-level data could be requested, as described in the example in Section 3.2.1, via the use of certain direct or indirect commands. Some restrictions as to the nature of the commands that can be made from the AC will therefore be necessary. In light of the findings from the previous chapter, it may also be necessary to mask the study-level data that arrives at the AC. As we have seen in Chapter 2, in certain circumstances some study-level data can, in principle, be identifying. Thus, particularly in the context of a GWAS, where large numbers of models are fitted and numerous results derived, perhaps only the “overall” results should ever be visible to the lead researchers.

A straightforward way of discouraging potential misuse of the DataSHIELD system is to create a log of all the instructions ever requested from an AC. These could be paired with unique researcher IDs, so the perpetrator of any potential malicious use of DataSHIELD can be identified. Other ways in which to uphold the security of DataSHIELD may include setting up a firewall around the DCs and the AC, and by encrypting the data shared between computer nodes.

3.3. Simulation Studies

This section illustrates the use of DataSHIELD in two different scenarios. Scenario 1 simulates a normally distributed outcome and demonstrates the fitting of a linear regression model in a DataSHIELD-type analysis – as outlined

in Section 3.2.3. In contrast, Scenario 2 simulates case-control data, and demonstrates an ILMA analysis for a logistic regression model in DataSHIELD – as outlined in Section 3.2.4. Note that this chapter demonstrates the validity of the mathematics behind DataSHIELD only. As has already been stated, work on the required IT system remains ongoing, and it is therefore not possible, at present, to fully demonstrate use of DataSHIELD in a setting where the data are truly distributed in different locations. The simulated data in the following examples are therefore held in a single location, and the procedures performed manually. In effect, the DataSHIELD analysis performed in Section 1 is therefore analogous to a conventional SLMA, because no transmission of the analysis requirements from the AC to the DCs, and vice-versa for the study level summary statistics, is actually performed (note that it is these transmissions that will ultimately provide DataSHIELD with improved flexibility compared to a conventional SLMA; however, this advantage cannot be achieved until work on the IT system is complete). In contrast, the procedure used in Scenario 2 cannot be considered a conventional SLMA, because it makes use of the DataSHIELD algorithm for fitting a GLM to distributed datasets. The DataSHIELD analysis performed in Scenario 2, thus, should produce identical results to an ILMA, while a conventional SLMA using a GLM would not.

This section is supported by Appendix C, which provides the corresponding R code and selected output from the fitting of the two illustrative models described here.

3.3.1 Scenario 1: Normally distributed data

This scenario simulates a hypothetical consortium of six studies set up to investigate determinants of blood pressure. The aim of the exercise is to demonstrate the use of DataSHIELD for an analysis using a linear model. Results are provided for two analyses: (1) an analysis of the complete, pooled dataset derived by combining the data from all studies (i.e. an ILMA); and (2) a DataSHIELD-type analysis (which, as stated above, is simply an SLMA here).

3.3.1.1 Simulation method

Six hypothetical studies are simulated to investigate the influence of age (AGE) and a single nucleotide polymorphism (SNP) on systolic blood pressure (SBP). The six studies consist of 1,000, 2,000, 3,000, 4,000, 2,500 and 2,500 participants respectively, with all participants aged between 50 and 70 years. For the j^{th} individual ($j = 1, \dots, 4,000$) in the i^{th} study ($i = 1, \dots, 6$), AGE_{ij} is generated from a uniform distribution with bounding parameters 50 and 70, and centred by subtracting the mean (60 years). SNP_{ij} is generated as the sum of two calls from a Bernoulli distribution with $p = 0.2$, corresponding to a minor-allele frequency of 0.2. The three genotypes are coded 0 (= no copies of the minor-allele), 1 (= one copy of the minor-allele) or 2 (= two copies of the minor-allele), reflecting an additive genetic model.

The linear predictor for each individual, LP_{ij} , is generated as:

$$LP_{ij} = B_{intercept} + B_{AGE} * AGE_{ij} + B_{SNP} * SNP_{ij},$$

where $B_{\text{intercept}} = 125$, $B_{\text{AGE}} = 0.5$, and $B_{\text{SNP}} = 0.3$. Subsequently, SBP_{ij} is generated (in mmHg) from a normal distribution with mean = LP_{ij} , and SD = 11.

3.3.1.2 Approach to Analysis

All analyses involve fitting the following linear regression model:

Equation 26

$$SBP_{ij} = b_{\text{intercept}} + b_{\text{AGE}} * AGE_{ij} + b_{\text{SNP}} * SNP_{ij} + \epsilon, \quad \epsilon \sim N(0, \sigma_{\epsilon}^2)$$

Analysis 1 applies Equation 26 to a dataset formed by pooling the data from all six studies (ignoring the subscript i and where $j = 1, \dots, 15,000$). Analysis 2 fits Equation 26 to each study individually, before combining the results using a weighted average. This analysis, hence, is analogous to the DataSHIELD analysis described in Section 3.2.3, in which each dataset is analysed on a separate DC before the AC synthesises the results by taking weighted averages. The data here remain stored on a single PC, however, and no use of a DataSHIELD-type IT infrastructure is employed.

Appendix C.1 contains the R code used to program the two analyses.

3.3.1.3 Results

Table 30 below presents the final results for both analyses, while interim results for the study-specific analyses (as performed for Analysis 2 only) are also listed in Appendix C.1.

Coefficient:	Analysis 1		Analysis 2	
	<i>Estimate</i>	<i>Standard Error</i>	<i>Estimate</i>	<i>Standard Error</i>
<i>Intercept</i>	125.1577	0.1094	125.1541	0.1094
<i>AGE</i>	0.2594	0.0155	0.2595	0.0155
<i>SNP</i>	0.4480	0.1581	0.4582	0.1580

Table 30: Results for an ILMA analysis (Analysis 1) and a DataSHIELD SLMA analysis (Analysis 2) of the simulated SBP data.

Table 30 shows that the two analyses yield almost identical results, both in terms of the regression coefficients and the standard errors. This DataSHIELD analysis is simply an SLMA, but the IT infrastructure that will ultimately be involved in its application will, in future, provide improved flexibility compared to a conventional SLMA of distributed datasets.

3.3.2 Scenario 2: Binary data

Scenario 2 simulates case-control studies with the aim of demonstrating the use of a logistic regression model in DataSHIELD. As before, two analyses are undertaken: (1) an ILMA of the combined dataset; and (2) a DataSHIELD-type analysis fitted to the distributed (or *horizontally-partitioned*) dataset.

3.3.2.1 Simulation Method

This scenario simulates six hypothetical case-control studies set up to investigate the relationship between the risk of acute myocardial infarction (MI), body mass index (BMI), and a single nucleotide polymorphism (SNP). For the

j^{th} individual in the i^{th} study, BMI_{ij} is generated from a normal distribution with mean 23 kg/m^2 and standard deviation 4 kg/m^2 , and then centred by subtracting the mean, 23 kg/m^2 , from each measurement. A genotype for the SNP of interest, SNP_{ij} , is generated for each individual in a manner equivalent to the sum of two calls to a Bernoulli distribution with $p = 0.3$. The minor-allele frequency is thus 0.3, and each genotype is either 0 (= no copies of the minor-allele), 1 (= one copy of the minor-allele) or 2 (= two copies of the minor-allele). Given the coding of the SNP variable, the simulated data reflect an additive genetic model.

In addition to the regression coefficients for the intercept ($b_{\text{intercept}}$) and two simulated covariates, b_{BMI} and b_{SNP} , the model also incorporates an interaction term, $b_{\text{BMI.456}}$, to allow for between-study heterogeneity in the magnitude of the effect of the BMI covariate on the log-odds of MI. The interaction covariate takes the value zero for individuals in studies 1, 2, and 3, and the BMI value for individuals in studies 4, 5, and 6, while the interaction coefficient $b_{\text{BMI.456}}$ implies that a one unit change in BMI in a subject in studies 4, 5 or 6 increases the log-odds of MI by an amount $b_{\text{BMI.456}}$ higher than an equivalent change in a subject in studies 1, 2 or 3.

The following model is thus used to generate the linear predictor:

$$LP_{ij} = b_{\text{intercept}} + b_{\text{BMI}} * BMI_{ij} + b_{\text{BMI.456}} * BMI.456_{ij} + b_{\text{SNP}} * SNP_{ij},$$

where $b_{\text{intercept}} = -0.3$, $b_{\text{BMI}} = 0.02$, $b_{\text{BMI.456}} = 0.04$, and $b_{\text{SNP}} = 0.5$.

Probabilities for developing acute myocardial infarction, p_{ij} , are derived by taking the inverse logistic (expit) transformation of the linear-predictor:

$$p_{ij} = \exp(LP_{ij})/[1 + \exp(LP_{ij})].$$

Case-control status, y_{ij} , is then generated for each individual by taking a random draw from a Bernoulli distribution with $p = p_{ij}$:

$$y_{ij} \sim \text{Bernoulli}(p_{ij}).$$

If the sampled value of y_{ij} is 1, the subject is designated to be a case; if y_{ij} is 0 the subject is designated a control.

The simulation code for this example is provided in Appendix C2. The case-control composition of the six simulated studies is summarised in Table 31 below.

Study	Cases	Controls	Total
1	962	1038	2000
2	1486	1514	3000
3	761	739	1500
4	143	157	300
5	1031	969	2000
6	357	343	700

Table 31: Numbers of cases and controls in the six simulated studies.

In order to mimic the conditions of a real DataSHIELD analysis as closely as possible, each dataset is saved to a separate file and folder. An additional, “AC” directory is also created in which a file containing current values of the regression coefficients is saved. This is a “communal” folder, which is both accessible and writable in both the DC and the AC stages of the analysis (see following section).

3.3.2.2 Approach to Analysis

Analysis 1 fits the following logistic regression model to a pooled dataset containing the individual-level data from the six simulated case-control studies:

Equation 27

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = b_{\text{intercept}} + b_{\text{BMI}} * BMI_{ij} + b_{\text{BMI.456}} * BMI.456_{ij} + b_{\text{SNP}} * SNP_{ij}.$$

In effect, the DataSHIELD analysis (Analysis 2) also fits the model in Equation 27, but to the data from each of the six studies separately. The data here can be considered to be horizontally-partitioned, however, and, hence, the algorithm described in Section 3.2.4.2 is used to perform the analysis.

Analysis 2 is undertaken in stages, and the results from each stage are presented in Appendix C2.4. Firstly, a “master process” (MP) is performed at the AC to save a vector of initial values for the regression coefficients into the “AC” directory. In this case, this vector simply specifies each parameter coefficient to be zero, i.e. $\mathbf{b}_1 = (0, 0, 0, 0)$.

The first iteration of the analysis then commences by performing a set of “slave processes” (SPs) from each DC. The initial values of the regression coefficients are used to derive the expected information matrix, $\mathbf{I}_i(\hat{\mathbf{b}}_1)$, the score vector, $\mathbf{s}_i(\hat{\mathbf{b}}_1)$, and the log-likelihood, $\log L_{C1i}$, for each study (following equations 22, 23, and 25 respectively), before these components are saved to the AC directory as a distinct file for each study.

A further MP is then performed at the AC to complete the first iteration. The six files containing the summary components from each study are loaded and the respective components summed, before updated values for the regression coefficients and standard errors are derived following Equation 21.

Finally, a test for convergence is undertaken by summing the six $\log L_{C1i}$ values and following the procedure described in Section 3.2.4.2. When the convergence criterion is not met, the current values of the regression coefficients, $\mathbf{b}_{r=2}$, are written to the corresponding file in the “AC” directory, and,

as shown in Appendix C2.4, and the procedure is restarted. These values are then used to derive $I_i(\hat{\mathbf{b}}_2)$, $s_i(\hat{\mathbf{b}}_2)$, and $\log L_{C2i}$, and so on until convergence is achieved. Upon convergence, all useful output (i.e. final estimates of the regression coefficients and standard errors, and the model deviance) is saved to an “output” file in the AC directory.

In addition to the step-by-step results for the above analyses, Appendix C2 also contains R code for performing these analyses.

3.3.2.3 Results

Results for Scenario 2 are shown in Table 32 below.

Coefficient:	Analysis 1: ILMA		Analysis 2: DataSHIELD	
	<i>Estimate</i>	<i>Std Error</i>	<i>Estimate</i>	<i>Std Error</i>
<i>Intercept</i>	-0.32956	0.02838	-0.32960	0.02838
<i>BMI</i>	0.02300	0.00621	0.02300	0.00621
<i>BMI.456</i>	0.04126	0.01140	0.04126	0.01140
<i>SNP</i>	0.55173	0.03295	0.55170	0.03295

Table 32: Results for an ILMA analysis (Analysis 1) and a DataSHIELD analysis (Analysis 2) of the six MI case-control studies.

As can be seen, apart from rounding errors, the two analyses yield identical results. DataSHIELD, thus, loses no information compared to an ILMA, despite pooling none of the individual-level data from the different studies. These results, furthermore, demonstrate that between-study heterogeneity can also easily be accounted for in these analyses, simply by including study-specific interaction terms (in this case, the term BMI.456).

3.4. Discussion

We have seen that DataSHIELD is a new approach to synthesising data between studies, comprising an infrastructure that allows for the central coordination and execution of analyses, while using statistical algorithms that ensure identical results to existing approaches.

For the analysis of linear models, the DataSHIELD infrastructure enhances the flexibility of a conventional SLMA (see Scenario 1), and provides a means by which to easily perform exploratory analyses (such as the fitting of interactions terms). For GLMs, DataSHIELD replicates the results of an ILMA (see Scenario 2), which is often considered the “gold-standard” approach to data synthesis (Sutton *et al.*, 2008; Riley, *et al.*, 2010).

Any analysis in DataSHIELD is performed in parallel on a network of DCs. As such, descriptive statistics and models can be fitted quickly, with potentially little lag compared to an ILMA. Each DC contains the individual-level data for a particular study, and shares only summary statistics with the AC. In principle, DataSHIELD thus circumvents many of the ethico-legal stipulations that can restrict data use.

3.4.1 Further ethico-legal issues

Although the DataSHIELD approach avoids sharing any individual-level data beyond each DC, some ethico-legal issues surrounding its use remain. For example, as has been mentioned in Section 3.2.5, it could be argued that because DataSHIELD provides a lead research group the capability to derive

descriptive statistics, its use could anyway be viewed as analogous to having full access to the individual-level data from each study.

Related to this point, another key issue concerns how to implement safeguards in DataSHIELD to prevent its potential misuse, for example, by allowing a lead researcher to retrieve individual-level data from a participating study. As demonstrated in Section 3.2.1, without appropriate restrictions to its use, someone in control of the AC would be capable of extracting individual-level data from a participating study simply by requesting particular summaries of the data. Similarly, potentially sensitive information could be inferred from particular requests made using DataSHIELD, such as the dates of birth (or other potentially identifiable information) for all cases in a particular study.

In order to gain ethico-legal approval for its use, a number of safeguards against the above issues will need to be implemented in DataSHIELD. The IT system, thus, needs to be designed with these dangers in mind. The following section discusses the potential requirements of this IT system further.

3.4.2 The IT system

The key benefit that DataSHIELD provides is that it allows data to be synthesised without sharing any individual-level data. However, unless appropriate restrictions to its use can be put in place, this benefit could be lost completely, if it were to allow individual-level data to be requested from the AC. The software wrapper ultimately used to implement DataSHIELD must therefore restrict the execution of certain commands.

One way of restricting potential requests for individual-level data in DataSHIELD, as suggested in Section 3.2.5, is to allow the derivation only of “overall” statistics, e.g. those obtained from a weighted average of summary statistics from all participating studies. This would be relatively straightforward to implement in the software wrapper by restricting the user interface to the performance only of pre-selected functions. Although this would be restrictive to the researchers in charge of the AC, a number of “approved” functions could still be performed – such as the fitting of different models and the derivation of certain descriptive statistics – and the system would remain advantageous compared to a conventional SLMA.

More sophisticated solutions to the problem could involve restricting the execution of particular commands only. For example, study-specific requests – such as for a summary statistic from a particular study – could be restricted. Similarly, any requests that relate to a single individual could be blocked from returning user viewable output. Alternatively, some variables could be masked completely (e.g. date of birth), or they could be converted to different, non-identifiable forms (e.g. date of birth could be stored, but only age in years made visible).

As suggested in Section 3.2.5, security around the DataSHIELD system should be upheld by the installation of firewalls around the DCs, and by the use of encryption around any data transmitted to and from the AC. Technology to implement these guards already exists and, in principle, should be straightforward to incorporate into the DataSHIELD system.

3.4.3 Further developments

Other than the IT system, further work is required to investigate the feasibility of using different model types in DataSHIELD. For instance, the usage of mixed-models is becoming increasingly common in science, so an extension of the mathematics underpinning DataSHIELD to fit a mixed-model would be of interest.

An extension of the DataSHIELD approach to handling “vertically-partitioned” data (Karr *et al.*, 2007), in which different attributes on the same individuals are distributed in different databases, would also be useful. Although this scenario moves beyond the realm of synthesising data between GWAS, which is the original focus of this chapter, existing epidemiological studies such as ALSPAC (Golding, 1996) do have links to other protected databases. As such, an extension of DataSHIELD to the synthesis of vertically-partitioned data could help to link these resources, which otherwise may not have the required permissions to share identifiable information beyond the approved staff at each site.

3.4.4 Conclusions

The view that DataSHIELD offers a feasible solution to the real issues regarding data sharing in genomics is supported by the existence of similar software that is already in use. For example, the Economic and Social Research Council Secure Data Service (ESRC_Secure_Data_Service, 2009) provides an interface that allows users to request specific queries to a database, and to extract summary results only. Thus, the main barrier to the successful

implementation of DataSHIELD will be finding ways around the ethico-legal issues concerning its use. Eventually, should workable restrictions to the use of DataSHIELD be put into place, then, from an ethico-legal perspective, DataSHIELD might be viewed along the same lines as conventional SLMAs. From that point, DataSHIELD could then begin to be introduced as an alternative to SLMA for the synthesis of data between studies in genomics.

Conclusions and Further Work

This thesis investigates methods for combining or inferring information from genetic epidemiological studies by considering three issues of current importance to the field. Chapter 1 focuses on how to incorporate treatment information in observational studies of blood pressure (BP), and, thus, investigates approaches to make efficient use of study resources, i.e. by maximising the statistical power. In contrast, Chapter 2 addresses the issue of participant privacy in GWAS. This chapter examines the validity of a new class of statistical methodology that potentially allows inferences regarding participant presence within genome-wide association studies (GWAS) to be made. Building upon this, Chapter 3 introduces a novel approach to combining data between studies, which potentially avoids infringing the ethico-legal stipulations that restrict data use, while maintaining the security and privacy of study data. The key conclusions from each chapter are now summarised and some areas for further work discussed.

Chapter 1

In Chapter 1, different approaches to handling data from individuals who use treatment are compared under different, realistic conditions, and recommendations are provided as to most appropriate approaches to use in practice.

In general, the results from this chapter support previous findings (White *et al.*, 1994; White *et al.*, 2003; Tobin *et al.*, 2005; McClelland *et al.*, 2008). Under conditions in which the intervention is *non-differential*, i.e. where both the effect of treatment and the chance of receiving treatment depend only on BP, the best approaches to analysis are the “Informative BP” approaches. These methods make use of all the observed information and apply simple corrections for treatment. They therefore typically yield high powers and return the correct type I error rates. When the intervention is *differential*, however, the Informative BP approaches may yield biased estimates of the effects of any parameters involved. This can impact upon the type I error rates of the approaches and it can reduce the statistical power. Real analyses therefore require caution in the interpretation of results – particularly because any particular genetic variants involved in a differential intervention will often be unknown.

A logical extension of this work is to further investigate strategies for identifying any genetic variants possibly involved in a differential intervention. For example, some of the approaches yield biased estimates in the presence of a differential treatment effect while other approaches are unaffected; it may thus be possible to infer possible variants that interact with treatment by cross-

checking results between the different approaches. Alternatively, although modelling treatment as a covariate is known to be inadvisable, it may be possible to explicitly test for pharmacogenetic interactions simply by modelling the SNP-treatment interaction term for a particular SNP. Although provisional findings suggest these approaches may be useful, further follow-up work is required to assess how useful they really are.

In order to better understand the full implications of this work, further work is also needed to clarify the effects of a differential intervention on variables that are either correlated or associated with a parameter that is directly involved. For instance, the assumption that individuals with diabetes receive antihypertensive medication at a lower threshold of BP than non-diabetics implies that the effect on BP of any genetic variant associated with diabetes is likely to be biased too. Further work is required to quantify these potential biases in different realistic settings. If the bias extends substantially to estimates of the effects of variables associated with or in linkage disequilibrium with a “differentiating parameter”, this may have implications for the interpretation of results from genome-wide association studies (GWAS). For example, any SNPs associated with diabetes may have spurious association with BP; it may therefore be necessary in future to cross-check the results from GWAS of BP with those from GWAS of diabetes.

Chapter 2

Chapter 2 examines a set of statistical approaches recently proposed that claim to be able to use allele frequencies for a large number of SNPs to test

probabilistically whether a given individual contributed DNA to a genomic mixture. Although the original “Homer” method (Homer *et al.*, 2008) is ultimately shown to be of limited use in practice, alternative tests – such as those proposed by Visscher *et al.* (Visscher *et al.*, 2009) – appear valid. Hence, it does seem possible, under ideal conditions, to reliably infer presence within GWAS consisting of several hundred or even a thousand or more participants.

Any attempts to identify are conditional upon adherence to the key assumption of co-ancestry. In the context of case-control GWAS, this assumption appears reasonable, as case-control GWAS are usually well matched in ancestry. Correlation between SNPs – or *LD* – is another factor that can cause problems unless appropriately handled. A GEE approach is proposed to adjust for LD, and, as long as only weak levels of LD are present, this seems to be effective. Nevertheless, highly correlated data pose further problems, because GEE approaches assume no between-“cluster” correlation. Further work is required to investigate how best to handle highly correlated SNPs. Although increasing the cluster size improves the ability to adjust for LD, the main problem seems to be dealing with any between-cluster correlation. As such, different approaches to arranging the SNPs into clusters need to be trialled, so that any LD between SNPs in different clusters is avoided.

If an appropriate strategy for handling highly correlated data can be found, huge numbers of SNPs (such as a million or more) could potentially be used in these tests. In principle, this would provide sufficient power to identify individuals in much larger studies than those assessed in this work. Many of the recent findings from GWAS have resulted from meta-analyses consisting of tens of

thousands of participants in total (e.g. (Zeggini *et al.*, 2008; Levy *et al.*, 2009; Newton-Cheh *et al.*, 2009)). Further work is required to investigate whether allele frequencies could also compromise participant identity in studies of this type. In addition to the large sample sizes involved in these studies, a further complication is that, typically, these meta-analyses include studies from multiple countries. The implications of these methods and, in particular, their reliance on the co-ancestry assumption therefore need further clarification for these situations.

A further issue arising from this work concerns the general use of GEE models, and providing better guidelines for their use. For example, fitting a GEE with an independence correlation structure with clusters containing single observations can provide markedly different estimates of the variance of the regression parameters to fitting a GEE model with an independence correlation structure and larger cluster sizes. The clustering of the data is, thus, of paramount importance in the fitting a GEE model, and this probably needs to be emphasised. Additional work must be undertaken to further investigate this issue, but this goes beyond the scope of this thesis.

Chapter 3

Chapter 3 proposes a new strategy to pooling data between studies – called DataSHIELD. DataSHIELD encompasses an IT infrastructure that allows for the central coordination of analyses while avoiding the need to share any individual-level data beyond the original researchers involved in each study. It, thus, provides a flexible approach to data synthesis that circumvents the ethico-

legal restrictions to the sharing of data. DataSHIELD, in addition, guarantees to produce identical results to an optimal “individual-level meta-analysis” for the class of statistical model known as *generalised linear models*.

The simulation studies in Chapter 3 illustrate the use of DataSHIELD in two scenarios – demonstrating the performance of a linear and a logistic model. A useful extension of this work, however, would be to investigate how to fit further classes of statistical models in DataSHIELD, such as random-effects and mixed models. In principle, these models could be fitted relatively easily by adopting a coordinated SLMA procedure similar to that used for the analysis of the normally distributed outcome in Scenario 1. However, this will not necessarily provide identical results to a corresponding ILMA fitted with each model type as we showed for the generalised linear model case. It may not be straightforward to derive a set of statistical algorithms to guarantee identical results to an ILMA for other types of model, and extensive follow-up work would be required to investigate this. In addition, further consideration of the nature of the information disclosed by different model types would be needed before implementing these models in DataSHIELD. For instance, mixed models often involve large numbers of model parameters, which, in some circumstances, could become identifying in the same way that allele frequencies can potentially disclose an individual’s presence in a study (as seen in Chapter 2). Further work would therefore be needed to decide how to handle the extra information that these models convey.

Related to the above issue, further consideration is also needed to decide how to handle the data transmitted from each study to the analysis centre. This

information may also be considered to be potentially “identifying” and, hence, must be handled securely. It may be necessary to limit access to these data or possibly to prevent access to these data altogether (and, thus, output only the final, “overall” results to the users).

The major area for further work on DataSHIELD concerns the development of the software wrapper. In function, this needs to automate the derivation and transmission of the appropriate matrix components from each data centre (DC) to the analysis centre (AC), while preventing access to any individual-level data from the AC. It also needs to ensure that any communications between centres remains secure, for example, by encrypting any data sent to and from the AC. While the above functions will require the expertise specifically of computer scientists or software engineers, more generally the software wrapper must also include restrictions to limit its use to the collection only of summary statistics from each study – rather than the acquittal of data that convey individual-level information. Precisely how best to do this requires further thought, although perhaps a simple deterrent to the misuse of DataSHIELD would be to create a log of all requests from the DC, in addition to a user-identifier.

Final Conclusions

Due to the varied nature of the work in this thesis, the overall implications of this work must also be considered on a chapter-by-chapter basis.

From Chapter 1, the findings can help to inform the strategy for analysis in future studies of BP. Largely the results in Chapter 1 suggest maintaining the status-quo in the choice of approach to analysis. However, some of the

problems highlighted in this chapter urge caution in the interpretation of particular results, and demonstrate conditions under which the results can become distorted.

Chapter 2 helps clarify the dangers associated with the release of large numbers of aggregate statistics from GWAS, such as allele frequencies. This work demonstrates that “signed” statistics can be informative of an individual’s presence in a study, and it advises on what can be published from GWAS in spite of these potential threats to participant confidentiality.

Chapter 3 describes a practical solution to the real problems associated with pooling data between studies. The mathematical properties of the algorithms used in DataSHIELD guarantee identical results to the ideal, *individual-level* meta-analysis, and the IT infrastructure involved offers clear advantages over the conventional *study-level* meta-analysis (SLMA). As such, DataSHIELD will begin to be piloted in real studies over the coming months ahead, with the hope that one day its use may replace SLMA as the method of choice for the synthesis of results from GWAS.

Appendix A.

The following pages present the full tables of results for the scenarios simulated in Chapter 1. Each table shows the mean estimate, SE, and 80% coverage interval of the parameter coefficients for each approach, as well as Monte Carlo estimates of the statistical power and type I error rate. Note that unless stated otherwise in the main text, the simulated values of the parameter coefficients are: Intercept = 110; AGE = 0.4; SEX = 3; g_1 (shown as *Gene* in the tables) = 2; g_2 (shown as *Gene2* in the tables) = 0. These values represent the values of the coefficients $\beta_0 - \beta_4$ respectively.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		110.0592	0.3986	0.0254	0.792	3.0191	0.8059	0.782	2.0201	0.6220	0.801	0.0140	0.6217	0.801	0.900	0.058
(a)	No Adjustment	111.1889	0.3090	0.0213	0.005	2.3478	0.6760	0.610	1.5707	0.5218	0.644	0.0159	0.5216	0.787	0.849	0.052
(b)	Exclude	109.4370	0.2929	0.0261	0.004	2.2364	0.8186	0.604	1.5054	0.6371	0.664	0.0192	0.6313	0.808	0.652	0.043
(c)	Treatment as binary covariate*	111.9432	0.2491	0.0209	0.000	1.8988	0.6480	0.340	1.2707	0.5000	0.423	0.0156	0.4993	0.799	0.708	0.054
(d)	Binary Phenotype**	-2.7237	0.0362	0.0032	0.000	0.2779	0.0966	0.000	0.1845	0.0741	0.000	0.0007	0.0745	0.811	0.697	0.043
(e)	Fixed Treat Effect (c = 5)	110.8110	0.3389	0.0224	0.082	2.5721	0.7106	0.716	1.7204	0.5484	0.733	0.0160	0.5482	0.797	0.889	0.050
	Fixed Treat Effect (c = 10)	110.4332	0.3688	0.0238	0.484	2.7963	0.7563	0.781	1.8700	0.5837	0.792	0.0162	0.5835	0.796	0.889	0.053
	Fixed Treat Effect (c = 15)	110.0553	0.3987	0.0255	0.790	3.0206	0.8114	0.791	2.0197	0.6262	0.799	0.0164	0.6260	0.794	0.897	0.055
(f)	Fixed Substitution (m=130)	113.6856	0.2285	0.0189	0.000	1.7293	0.5987	0.215	1.1470	0.4620	0.284	0.0099	0.4619	0.805	0.694	0.045
	Fixed Substitution (m=140)	112.9299	0.2883	0.0202	0.000	2.1778	0.6422	0.510	1.4463	0.4956	0.552	0.0102	0.4955	0.797	0.831	0.043
(g)	Random Substitution	112.1833	0.3480	0.0235	0.170	2.6315	0.7458	0.722	1.7394	0.5756	0.756	0.0085	0.5754	0.789	0.865	0.042
(h)	Median Method (k = 140)	115.9177	0.2922	0.0203	0.000	1.5066	0.6378	0.150	1.0158	0.4493	0.178	0.00932	0.45429	0.828	0.465	0.013
	Median Method (k = 150)	107.5213	0.4545	0.0377	0.422	3.3657	1.1845	0.778	2.2193	0.8617	0.801	0.03429	0.8759	0.809	0.603	0.016
	Median Method (k = 160)	105.6733	0.4875	0.0586	0.429	3.7950	1.3785	0.753	2.7111	1.0876	0.729	0.04578	1.04787	0.820	0.546	0.014
(i)	Non-parametric Adjustment	110.2718	0.4065	0.0260	0.752	3.0756	0.8269	0.789	2.0529	0.6382	0.805	0.0142	0.6380	0.797	0.895	0.047
(j)	Censored Normal Regression	108.4322	0.4244	0.0275	0.632	3.2247	0.8673	0.764	2.1659	0.6718	0.785	0.0190	0.6690	0.798	0.899	0.046

Table 33: Results for 1,000 runs of the General Simulation Study.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach (c) also yields a mean treatment coefficient of 10.0154, with a mean standard error of 0.7404.

** Parameter estimates for Approach (d) are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		109.9524	0.4008	0.0260	0.787	2.9796	0.8257	0.781	2.0143	0.6373	0.808	0.0051	0.6375	0.788	0.886	0.051
(a)	No Adjustment	110.9948	0.3123	0.0218	0.004	2.3231	0.6938	0.610	1.5808	0.5355	0.668	-0.0061	0.5357	0.797	0.843	0.056
(b)	Exclude	109.1820	0.2942	0.0267	0.005	2.1846	0.8378	0.606	1.4950	0.6520	0.669	0.0023	0.6466	0.789	0.632	0.055
(c)	Treatment as binary covariate*	111.7326	0.2495	0.0213	0.000	1.8592	0.6627	0.321	1.2728	0.5113	0.429	-0.0117	0.5110	0.798	0.701	0.062
(d)	Binary Phenotype **	-2.6672	0.0356	0.0031	0.000	0.2647	0.0964	0.000	0.1757	0.0739	0.000	0.0016	0.0743	0.813	0.662	0.048
(e)	Fixed Treat Effect (c = 5)	110.6474	0.3418	0.0230	0.114	2.5413	0.7294	0.702	1.7257	0.5629	0.741	-0.0037	0.5631	0.792	0.869	0.058
	Fixed Treat Effect (c = 10)	110.3001	0.3713	0.0244	0.545	2.7595	0.7757	0.755	1.8706	0.5987	0.796	-0.0013	0.5989	0.785	0.881	0.055
	Fixed Treat Effect (c = 15)	109.9528	0.4009	0.0262	0.782	2.9778	0.8311	0.776	2.0155	0.6414	0.810	0.0011	0.6417	0.791	0.889	0.051
(f)	Fixed Substitution (m=130)	113.4538	0.2304	0.0193	0.000	1.6989	0.6116	0.203	1.1445	0.4720	0.301	-0.0017	0.4722	0.784	0.688	0.059
	Fixed Substitution (m=140)	112.7591	0.2895	0.0206	0.000	2.1354	0.6559	0.469	1.4343	0.5062	0.558	0.0031	0.5064	0.784	0.804	0.056
(g)	Random Substitution	112.0609	0.3486	0.0239	0.196	2.5728	0.7591	0.727	1.7239	0.5859	0.760	0.0065	0.5861	0.786	0.828	0.054
(h)	Median Method (k = 160)	105.6368	0.4866	0.0598	0.451	3.8194	1.3986	0.738	2.6094	1.1020	0.762	0.0692	1.0596	0.801	0.512	0.029
	Median Method (k = 180)	105.6127	0.4870	0.0624	0.465	3.8232	1.4210	0.746	2.6162	1.1329	0.772	0.0701	1.0757	0.814	0.499	0.024
	Median Method (k = 200)	105.6130	0.4870	0.0621	0.460	3.8234	1.4226	0.747	2.6163	1.1349	0.764	0.0694	1.0766	0.808	0.492	0.024
(i)	Non-parametric Adjustment	110.1375	0.4124	0.0269	0.734	3.0553	0.8546	0.783	2.0627	0.6596	0.805	0.0047	0.6598	0.791	0.880	0.048
(j)	Censored Normal Regression	108.2769	0.4291	0.0283	0.586	3.1960	0.8928	0.772	2.1674	0.6914	0.793	0.0083	0.6892	0.795	0.881	0.051

Table 34: Results for 1,000 runs of Scenario 1.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach (c) also yields a mean treatment coefficient of 10.6380, with a mean standard error of 0.7556.

** Parameter estimates for Approach (d) are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		109.9574	0.4014	0.0253	0.795	2.9908	0.8050	0.789	1.9898	0.6213	0.809	-0.0264	0.6208	0.805	0.896	0.057
(a)	No Adjustment	109.2213	0.3411	0.0235	0.109	2.5489	0.7457	0.708	1.7087	0.5756	0.735	-0.0240	0.5751	0.805	0.844	0.052
(b)	Exclude	108.8349	0.3432	0.0280	0.232	2.5444	0.8838	0.722	1.7417	0.6858	0.761	-0.0243	0.6817	0.812	0.704	0.047
(c)	Treatment as binary covariate*	109.1751	0.3374	0.0237	0.095	2.5218	0.7460	0.699	1.6918	0.5757	0.724	-0.0234	0.5750	0.808	0.838	0.052
(d)	Binary Phenotype **	-2.1039	0.0308	0.0030	0.000	0.2296	0.0933	0.000	0.1521	0.0718	0.000	-0.0030	0.0719	0.817	0.566	0.040
(e)	Fixed Treat Effect (c = 5)	109.4672	0.3612	0.0238	0.357	2.6952	0.7548	0.747	1.8033	0.5826	0.782	-0.0244	0.5821	0.804	0.876	0.055
	Fixed Treat Effect (c = 10)	109.7130	0.3812	0.0244	0.670	2.8415	0.7768	0.770	1.8979	0.5996	0.799	-0.0248	0.5991	0.813	0.891	0.055
	Fixed Treat Effect (c = 15)	109.9589	0.4013	0.0255	0.795	2.9878	0.8107	0.790	1.9925	0.6257	0.811	-0.0251	0.6252	0.809	0.895	0.052
(f)	Fixed Substitution (m=130)	114.3950	0.2473	0.0202	0.000	1.8263	0.6410	0.293	1.2363	0.4947	0.401	-0.0236	0.4943	0.816	0.699	0.050
	Fixed Substitution (m=140)	114.8868	0.2874	0.0211	0.000	2.1189	0.6717	0.482	1.4255	0.5185	0.563	-0.0243	0.5180	0.827	0.777	0.045
(g)	Random Substitution	115.3666	0.3277	0.0241	0.051	2.4123	0.7646	0.654	1.6195	0.5902	0.723	-0.0209	0.5897	0.821	0.790	0.048
(h)	Median Method (k = 160)	107.4183	0.5107	0.0548	0.256	3.8260	1.5177	0.728	2.5947	1.1696	0.758	-0.0312	1.1554	0.818	0.460	0.023
	Median Method (k = 180)	107.3029	0.5126	0.0592	0.273	3.8521	1.5531	0.734	2.6385	1.2408	0.769	-0.0270	1.1908	0.830	0.415	0.021
	Median Method (k = 200)	107.3029	0.5127	0.0592	0.273	3.8506	1.5565	0.730	2.6389	1.2403	0.776	-0.0277	1.1918	0.822	0.427	0.022
(i)	Non-parametric Adjustment	112.6618	0.3787	0.0249	0.631	2.8076	0.7916	0.776	1.8778	0.6110	0.800	-0.0274	0.6105	0.803	0.861	0.048
(j)	Censored Normal Regression	109.7253	0.4215	0.0278	0.675	3.1346	0.8806	0.781	2.1051	0.6813	0.812	-0.0241	0.6791	0.811	0.862	0.049

Table 35: Results for 1,000 runs of Scenario 2.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach (c) also yields a mean treatment coefficient of 0.9271, with a mean standard error of 0.8344.

** Parameter estimates for Approach (d) are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		110.0195	0.4000	0.0253	0.805	2.9527	0.8052	0.797	1.9672	0.6213	0.796	-0.0013	0.6213	0.794	0.877	0.055
(a)	No Adjustment	112.5897	0.2595	0.0197	0.000	1.9232	0.6246	0.326	1.2878	0.4819	0.427	0.0022	0.4819	0.796	0.757	0.063
(b)	Exclude	109.4051	0.2939	0.0260	0.006	2.1888	0.8174	0.621	1.4849	0.6358	0.673	0.0097	0.6302	0.788	0.652	0.065
(c)	Treatment as binary covariate*	112.9375	0.2321	0.0199	0.000	1.7263	0.6192	0.207	1.1581	0.4775	0.313	0.0034	0.4771	0.793	0.679	0.063
(d)	Binary Phenotype**	-2.7395	0.0366	0.0032	0.000	0.2696	0.0967	0.000	0.1787	0.0742	0.000	-0.0010	0.0746	0.793	0.666	0.047
(e)	Fixed Treat Effect (c = 5)	112.2040	0.2897	0.0203	0.000	2.1396	0.6458	0.485	1.4309	0.4983	0.548	0.0007	0.4983	0.793	0.812	0.059
	Fixed Treat Effect (c = 10)	111.8183	0.3198	0.0214	0.010	2.3560	0.6806	0.632	1.5739	0.5251	0.671	-0.0007	0.5251	0.785	0.848	0.050
	Fixed Treat Effect (c = 15)	111.4326	0.3500	0.0229	0.172	2.5725	0.7269	0.740	1.7170	0.5608	0.752	-0.0022	0.5608	0.785	0.861	0.053
(f)	Fixed Substitution (m=130)	113.6216	0.2298	0.0188	0.000	1.6959	0.5983	0.172	1.1314	0.4616	0.270	0.0070	0.4616	0.787	0.700	0.056
	Fixed Substitution (m=140)	112.8502	0.2901	0.0202	0.000	2.1288	0.6418	0.483	1.4175	0.4951	0.558	0.0041	0.4952	0.790	0.816	0.058
(g)	Random Substitution	112.0779	0.3505	0.0235	0.194	2.5580	0.7450	0.733	1.7032	0.5748	0.744	0.0014	0.5748	0.793	0.838	0.058
(h)	Median Method (k = 160)	105.7912	0.4853	0.0582	0.435	3.7862	1.3725	0.740	2.5839	1.0760	0.747	0.0156	1.0332	0.795	0.530	0.016
	Median Method (k = 180)	105.7744	0.4856	0.0607	0.440	3.7886	1.3852	0.751	2.5912	1.1106	0.762	0.0145	1.0451	0.809	0.508	0.017
	Median Method (k = 200)	105.7737	0.4856	0.0604	0.443	3.7894	1.3858	0.747	2.5916	1.1095	0.762	0.0146	1.0458	0.808	0.518	0.015
(i)	Non-parametric Adjustment	111.6526	0.3417	0.0224	0.090	2.5122	0.7104	0.706	1.6755	0.5481	0.736	-0.0011	0.5481	0.787	0.863	0.054
(j)	Censored Normal Regression	109.6128	0.3775	0.0251	0.635	2.7885	0.7897	0.785	1.8728	0.6118	0.786	0.0003	0.6091	0.797	0.864	0.051

Table 36: Results for 1,000 runs of Scenario 3.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach C also yields a mean treatment coefficient of 4.5565, with a mean standard error of 0.7082.

** Parameter estimates for Approach D are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		110.0402	0.3992	0.0253	0.794	2.9954	0.8052	0.811	2.0302	0.6209	0.822	-0.0218	0.6213	0.793	0.919	0.047
(a)	No Adjustment	110.7321	0.3514	0.0228	0.202	2.6438	0.7254	0.741	1.7838	0.5594	0.787	-0.0272	0.5598	0.793	0.912	0.045
(b)	Exclude	109.4729	0.2931	0.0260	0.002	2.2275	0.8177	0.622	1.4953	0.6362	0.666	-0.0552	0.6309	0.789	0.641	0.045
(c)	Treatment as binary covariate*	112.0626	0.2475	0.0207	0.000	1.8827	0.6424	0.322	1.2613	0.4953	0.408	-0.0366	0.4951	0.789	0.721	0.049
(d)	Binary Phenotype**	-2.7260	0.0363	0.0032	0.000	0.2721	0.0967	0.000	0.1854	0.0741	0.000	-0.0003	0.0745	0.788	0.704	0.039
(e)	Fixed Treat Effect (c = 5)	110.3472	0.3815	0.0244	0.681	2.8636	0.7760	0.794	1.9345	0.5984	0.818	-0.0245	0.5988	0.798	0.913	0.041
	Fixed Treat Effect (c = 10)	109.9622	0.4115	0.0263	0.737	3.0833	0.8349	0.812	2.0853	0.6439	0.824	-0.0218	0.6443	0.808	0.915	0.039
	Fixed Treat Effect (c = 15)	109.5773	0.4415	0.0283	0.425	3.3031	0.9007	0.777	2.2360	0.6946	0.795	-0.0191	0.6951	0.798	0.908	0.040
(f)	Fixed Substitution (m=130)	113.6965	0.2288	0.0188	0.000	1.7175	0.5981	0.191	1.1451	0.4612	0.268	-0.0424	0.4615	0.795	0.691	0.045
	Fixed Substitution (m=140)	112.9266	0.2889	0.0202	0.000	2.1571	0.6417	0.481	1.4466	0.4948	0.549	-0.0371	0.4952	0.802	0.847	0.047
(g)	Random Substitution	112.1550	0.3490	0.0234	0.195	2.5995	0.7451	0.749	1.7469	0.5746	0.767	-0.0325	0.5750	0.816	0.864	0.044
(h)	Median Method (k = 160)	105.8509	0.4828	0.0582	0.481	3.8692	1.3629	0.724	2.6451	1.0819	0.734	0.0288	1.0424	0.817	0.543	0.017
	Median Method (k = 180)	105.8316	0.4831	0.0603	0.481	3.8753	1.3825	0.729	2.6536	1.1108	0.745	0.0269	1.0496	0.826	0.522	0.014
	Median Method (k = 200)	105.8309	0.4831	0.0604	0.481	3.8747	1.3791	0.730	2.6556	1.1141	0.748	0.0276	1.0518	0.826	0.519	0.015
(i)	Non-parametric Adjustment	109.7879	0.4625	0.0302	0.231	3.4466	0.9595	0.755	2.3329	0.7400	0.772	-0.0183	0.7405	0.790	0.894	0.039
(j)	Censored Normal Regression	108.0045	0.4656	0.0303	0.191	3.4940	0.9550	0.749	2.3716	0.7389	0.758	-0.0205	0.7369	0.802	0.905	0.040

Table 37: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 5%.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach C also yields a mean treatment coefficient of 17.3203, with a mean standard error of 0.7342.

** Parameter estimates for Approach D are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		109.9053	0.4016	0.0254	0.803	3.0045	0.8055	0.802	2.0101	0.6217	0.809	0.0026	0.6214	0.801	0.902	0.052
(a)	No Adjustment	111.3061	0.3057	0.0210	0.001	2.2929	0.6658	0.588	1.5415	0.5138	0.629	0.0054	0.5136	0.801	0.856	0.050
(b)	Exclude	109.3549	0.2943	0.0261	0.006	2.2222	0.8184	0.622	1.5112	0.6369	0.669	0.0191	0.6310	0.793	0.667	0.051
(c)	Treatment as binary covariate*	112.0687	0.2470	0.0206	0.000	1.8610	0.6392	0.310	1.2560	0.4931	0.417	0.0094	0.4925	0.802	0.715	0.047
(d)	<i>Binary Phenotype</i> **	-2.7406	0.0366	0.0032	0.000	0.2761	0.0967	0.000	0.1835	0.0742	0.000	-0.0031	0.0746	0.801	0.706	0.059
(e)	Fixed Treat Effect (c = 5)	110.9118	0.3359	0.0220	0.055	2.5154	0.6999	0.702	1.6888	0.5402	0.729	0.0037	0.5400	0.811	0.884	0.052
	Fixed Treat Effect (c = 10)	110.5175	0.3662	0.0235	0.434	2.7379	0.7454	0.764	1.8360	0.5753	0.785	0.0020	0.5751	0.806	0.902	0.050
	Fixed Treat Effect (c = 15)	110.1232	0.3964	0.0252	0.787	2.9605	0.8004	0.795	1.9832	0.6177	0.819	0.0003	0.6174	0.800	0.903	0.047
(f)	Fixed Substitution (m=130)	113.6087	0.2298	0.0188	0.000	1.7176	0.5987	0.190	1.1444	0.4621	0.281	0.0102	0.4619	0.788	0.693	0.051
	Fixed Substitution (m=140)	112.8201	0.2903	0.0202	0.001	2.1627	0.6421	0.496	1.4389	0.4956	0.544	0.0068	0.4954	0.807	0.822	0.054
(g)	Random Substitution	112.0276	0.3508	0.0235	0.211	2.6151	0.7453	0.733	1.7355	0.5752	0.753	0.0031	0.5750	0.807	0.847	0.055
(h)	Median Method (k = 160)	105.5650	0.4892	0.0583	0.426	3.8893	1.3708	0.701	2.6374	1.0767	0.735	-0.0034	1.0364	0.796	0.543	0.021
	Median Method (k = 180)	105.5377	0.4897	0.0607	0.426	3.8964	1.3879	0.707	2.6460	1.1100	0.753	-0.0046	1.0474	0.803	0.523	0.020
	Median Method (k = 200)	105.5382	0.4897	0.0607	0.431	3.8947	1.3907	0.701	2.6469	1.1078	0.759	-0.0054	1.0469	0.800	0.525	0.025
(i)	Non-parametric Adjustment	110.3519	0.4007	0.0255	0.783	2.9867	0.8092	0.793	1.9967	0.6245	0.801	-0.0003	0.6243	0.809	0.898	0.052
(j)	Censored Normal Regression	108.4963	0.4211	0.0271	0.659	3.1564	0.8554	0.787	2.1241	0.6626	0.796	0.0036	0.6597	0.809	0.900	0.049

Table 38: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 10%.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach C also yields a mean treatment coefficient of 9.7036, with a mean standard error of 0.7309.

** Parameter estimates for Approach D are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		110.0000	0.4007	0.0254	0.801	3.0025	0.8059	0.805	1.9573	0.6214	0.817	-0.0054	0.6218	0.787	0.880	0.057
(a)	No Adjustment	112.0746	0.2575	0.0200	0.000	1.9366	0.6360	0.345	1.2456	0.4904	0.409	0.0077	0.4907	0.801	0.697	0.056
(b)	Exclude	109.4049	0.2945	0.0261	0.003	2.2084	0.8187	0.606	1.4359	0.6368	0.633	0.0315	0.6316	0.813	0.619	0.052
(c)	Treatment as binary covariate*	112.2319	0.2451	0.0205	0.000	1.8447	0.6357	0.286	1.1844	0.4900	0.367	0.0095	0.4899	0.803	0.666	0.056
(d)	Binary Phenotype**	-2.7376	0.0366	0.0032	0.000	0.2719	0.0967	0.000	0.1802	0.0741	0.000	-0.0012	0.0745	0.792	0.679	0.059
(e)	Fixed Treat Effect (c = 5)	111.6850	0.2876	0.0205	0.000	2.1594	0.6496	0.481	1.3948	0.5009	0.528	0.0048	0.5012	0.798	0.789	0.058
	Fixed Treat Effect (c = 10)	111.2954	0.3178	0.0213	0.009	2.3821	0.6772	0.641	1.5440	0.5222	0.640	0.0020	0.5225	0.796	0.838	0.058
	Fixed Treat Effect (c = 15)	110.9058	0.3479	0.0226	0.157	2.6048	0.7171	0.742	1.6931	0.5530	0.729	-0.0009	0.5534	0.802	0.862	0.063
(f)	Fixed Substitution (m=130)	113.6543	0.2297	0.0189	0.000	1.7077	0.5987	0.203	1.0971	0.4617	0.244	0.0160	0.4620	0.800	0.653	0.054
	Fixed Substitution (m=140)	112.8751	0.2900	0.0202	0.000	2.1531	0.6421	0.468	1.3955	0.4951	0.519	0.0102	0.4955	0.795	0.809	0.058
(g)	Random Substitution	112.0875	0.3503	0.0235	0.197	2.5984	0.7452	0.749	1.6963	0.5746	0.743	0.0092	0.5750	0.794	0.840	0.064
(h)	Median Method (k = 160)	105.6804	0.4881	0.0585	0.407	3.8568	1.3686	0.733	2.5797	1.0806	0.745	0.0010	1.0410	0.812	0.520	0.027
	Median Method (k = 180)	105.6631	0.4884	0.0608	0.417	3.8603	1.3891	0.735	2.5846	1.1108	0.761	0.0000	1.0522	0.815	0.488	0.025
	Median Method (k = 200)	105.6624	0.4884	0.0609	0.416	3.8607	1.3893	0.730	2.5853	1.1149	0.771	-0.0002	1.0506	0.816	0.504	0.022
(i)	Non-parametric Adjustment	111.2264	0.3443	0.0223	0.126	2.5734	0.7076	0.718	1.6746	0.5456	0.720	0.0009	0.5460	0.797	0.861	0.064
(j)	Censored Normal Regression	109.0819	0.3800	0.0250	0.656	2.8501	0.7876	0.785	1.8657	0.6099	0.791	0.0079	0.6077	0.793	0.859	0.066

Table 39: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 15%.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach C also yields a mean treatment coefficient of 2.0654, with a mean standard error of 0.7266.

** Parameter estimates for Approach D are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis		Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP		110.0681	0.3991	0.0254	0.786	3.0343	0.8053	0.790	1.9973	0.6215	0.837	-0.0114	0.6215	0.810	0.915	0.039
(a)	No Adjustment	112.7637	0.2088	0.0201	0.000	1.5960	0.6384	0.188	1.0518	0.4927	0.260	0.0277	0.4927	0.794	0.569	0.044
(b)	Exclude	109.4686	0.2925	0.0261	0.003	2.2354	0.8181	0.606	1.4898	0.6366	0.656	0.0195	0.6307	0.793	0.644	0.043
(c)	Treatment as binary covariate*	112.3362	0.2424	0.0203	0.000	1.8492	0.6300	0.285	1.2182	0.4860	0.359	0.0209	0.4856	0.795	0.704	0.043
(d)	Binary Phenotype**	-2.7220	0.0363	0.0032	0.000	0.2789	0.0966	0.000	0.1834	0.0741	0.000	-0.0050	0.0745	0.802	0.687	0.046
(e)	Fixed Treat Effect (c = 5)	112.3843	0.2388	0.0198	0.000	1.8224	0.6290	0.278	1.2005	0.4855	0.343	0.0215	0.4855	0.795	0.694	0.042
	Fixed Treat Effect (c = 10)	112.0049	0.2689	0.0200	0.000	2.0487	0.6348	0.406	1.3493	0.4899	0.479	0.0153	0.4900	0.807	0.789	0.042
	Fixed Treat Effect (c = 15)	111.6255	0.2989	0.0206	0.000	2.2751	0.6553	0.558	1.4980	0.5058	0.606	0.0092	0.5058	0.814	0.852	0.039
(f)	Fixed Substitution (m=130)	113.7086	0.2281	0.0188	0.000	1.7341	0.5979	0.202	1.1355	0.4615	0.275	0.0130	0.4615	0.797	0.697	0.042
	Fixed Substitution (m=140)	112.9499	0.2882	0.0202	0.000	2.1868	0.6416	0.477	1.4330	0.4951	0.538	0.0006	0.4952	0.802	0.845	0.042
(g)	Random Substitution	112.1887	0.3484	0.0235	0.183	2.6344	0.7451	0.732	1.7277	0.5751	0.768	-0.0124	0.5751	0.823	0.871	0.041
(h)	Median Method (k = 160)	105.8662	0.4841	0.0581	0.411	3.9047	1.3694	0.716	2.6340	1.0758	0.740	-0.0358	1.0404	0.812	0.552	0.014
	Median Method (k = 180)	105.8522	0.4843	0.0606	0.424	3.9094	1.3883	0.716	2.6370	1.1078	0.759	-0.0366	1.0524	0.826	0.525	0.010
	Median Method (k = 200)	105.8517	0.4844	0.0605	0.434	3.9087	1.3856	0.717	2.6368	1.1075	0.759	-0.0371	1.0511	0.835	0.531	0.009
(i)	Non-parametric Adjustment	112.0886	0.2968	0.0203	0.000	2.2576	0.6445	0.541	1.4860	0.4974	0.583	0.0056	0.4975	0.807	0.864	0.039
(j)	Censored Normal Regression	109.4746	0.3458	0.0239	0.178	2.6391	0.7510	0.737	1.7508	0.5825	0.769	0.0046	0.5792	0.808	0.879	0.032

Table 40: Results for 1,000 runs of Scenario 4 with a proportional treatment effect of 20%.

Included are mean parameter estimates and S.E.'s, coverage rates based on 80% C.I.'s, and power and type I error relative to gene and gene2 respectively (at the 5% level of sig.).

* Approach C also yields a mean treatment coefficient of -5.5901, with a mean standard error of 0.7196.

** Parameter estimates for Approach D are log odds-ratios – and are therefore non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	109.9075	0.4017	0.0254	0.792	3.0092	0.8058	0.805	1.9788	0.6219	0.812	0.895
(a) No Adjustment	110.3273	0.3101	0.0213	0.005	2.3384	0.6769	0.608	2.8487	0.5224	0.338	0.999
(b) Exclude	109.2942	0.2954	0.0261	0.009	2.2417	0.8186	0.628	1.4868	0.6369	0.679	0.661
(c) Treatment as binary covariate*	111.0891	0.2506	0.0209	0.000	1.9052	0.6498	0.352	2.5618	0.5013	0.544	1.000
inc. Gene-Treat. Interaction**	111.8396	0.2495	0.0209	0.000	1.9281	0.6482	0.364	1.4557	0.5933	0.615	0.670
(d) <i>Binary Phenotype</i> ***	-2.7346	0.0366	0.0032	.	0.2706	0.0967	.	0.1796	0.0741	.	0.692
Fixed Treat Effect (c = 5)	109.9404	0.3403	0.0224	0.089	2.5582	0.7110	0.732	2.9938	0.5487	0.285	0.999
(e) Fixed Treat Effect (c = 10)	109.5535	0.3704	0.0238	0.493	2.7779	0.7563	0.787	3.1389	0.5837	0.247	0.999
Fixed Treat Effect (c = 15)	109.1667	0.4006	0.0255	0.794	2.9977	0.8110	0.809	3.2840	0.6259	0.223	0.999
(f) Fixed Substitution (m=130)	113.5562	0.2308	0.0189	0.000	1.7261	0.5988	0.197	1.1306	0.4621	0.276	0.694
Fixed Substitution (m=140)	112.7824	0.2911	0.0202	0.001	2.1656	0.6424	0.484	1.4208	0.4958	0.557	0.823
(g) Random Substitution	112.0094	0.3513	0.0235	0.201	2.6058	0.7457	0.747	1.7111	0.5755	0.761	0.855
Median Method (k = 160)	105.5431	0.4897	0.0582	0.419	3.8747	1.3709	0.745	2.6067	1.0743	0.732	0.549
(h) Median Method (k = 180)	105.5237	0.4900	0.0607	0.419	3.8798	1.3862	0.748	2.6128	1.1073	0.741	0.512
Median Method (k = 200)	105.5244	0.4900	0.0605	0.424	3.8773	1.3878	0.748	2.6132	1.1074	0.750	0.522
(i) Non-parametric Adjustment	109.5147	0.4102	0.0261	0.746	3.0583	0.8290	0.804	3.0543	0.6398	0.363	0.994
(j) Censored Normal Regression	107.7339	0.4272	0.0275	0.615	3.2068	0.8676	0.798	3.0555	0.6700	0.373	0.991

Table 41: Results for 1,000 runs of Scenario 5a when $\beta_3 = 2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 13.5531, with a mean standard error of 0.7299.

**Approach C with the gene-treatment interaction term yields a mean treatment coefficient of 10.5038 with mean standard error of 1.2686; and mean gene-treatment interaction of 2.9862 with mean standard error of 1.0067.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	112.2823	0.4024	0.0254	0.800	2.9432	0.8060	0.811	-2.0083	0.6221	0.776	0.888
(a) No Adjustment	112.1878	0.3095	0.0213	0.002	2.2817	0.6773	0.587	-0.3456	0.5228	0.036	0.120
(b) Exclude	111.1042	0.2949	0.0261	0.002	2.1990	0.8182	0.601	-1.4669	0.6262	0.642	0.649
(c) Treatment as binary covariate*	112.6031	0.2512	0.0210	0.000	1.8709	0.6516	0.322	-0.0576	0.5029	0.007	0.063
inc. Gene-Treat. Interaction**	113.5263	0.2495	0.0208	0.000	1.8995	0.6472	0.35	-1.4379	0.5829	0.596	0.681
(d) <i>Binary Phenotype</i> ***	-2.5183	0.0366	0.0032	.	0.2665	0.0967	.	-0.1876	0.0754	.	0.693
Fixed Treat Effect (c = 5)	111.9710	0.3398	0.0224	0.090	2.4956	0.7107	0.706	-0.4956	0.5485	0.085	0.162
(e) Fixed Treat Effect (c = 10)	111.7541	0.3701	0.0238	0.503	2.7094	0.7554	0.785	-0.6457	0.5830	0.152	0.212
Fixed Treat Effect (c = 15)	111.5373	0.4004	0.0255	0.801	2.9232	0.8094	0.806	-0.7957	0.6247	0.262	0.256
(f) Fixed Substitution (m=130)	114.9582	0.2303	0.0188	0.000	1.6968	0.5986	0.184	-1.1545	0.4620	0.288	0.705
Fixed Substitution (m=140)	114.5245	0.2909	0.0202	0.000	2.1244	0.6420	0.479	-1.4546	0.4955	0.570	0.817
(g) Random Substitution	114.0782	0.3517	0.0235	0.225	2.5463	0.7451	0.729	-1.7485	0.5751	0.741	0.851
Median Method (k = 160)	108.6090	0.4904	0.0587	0.407	3.7814	1.3806	0.737	-2.5348	1.0247	0.747	0.575
(h) Median Method (k = 180)	108.6045	0.4905	0.0608	0.407	3.7836	1.3883	0.737	-2.5369	1.0351	0.752	0.557
Median Method (k = 200)	108.6037	0.4905	0.0607	0.426	3.7848	1.3848	0.748	-2.5364	1.0344	0.754	0.558
(i) Non-parametric Adjustment	111.9571	0.4081	0.0259	0.759	2.9680	0.8230	0.808	-1.0837	0.6352	0.429	0.397
(j) Censored Normal Regression	110.2580	0.4274	0.0275	0.587	3.1295	0.8659	0.796	-1.2686	0.6650	0.551	0.468

Table 42: Results for 1,000 runs of Scenario 5a when $\beta_3 = -2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 14.8123, with a mean standard error of 0.7885.

**The additional analysis for (c) which includes the gene-treatment interaction term yields a mean treatment coefficient of 10.4919 with mean standard error of 1.3038; and mean gene-treatment interaction of 4.5431 with mean standard error of 1.0868.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Type I Error
Underlying SBP	111.9452	0.4010	0.0254	0.803	2.9813	0.8056	0.784	-0.0040	0.6215	0.797	0.054
(a) No Adjustment	111.9881	0.3075	0.0213	0.000	2.2899	0.6756	0.578	1.3047	0.5212	0.114	0.707
(b) Exclude	110.7932	0.2953	0.0263	0.003	2.2108	0.8259	0.605	0.0020	0.6369	0.823	0.041
(c) Treatment as binary covariate*	112.5053	0.2493	0.0209	0.000	1.8644	0.6500	0.319	1.3019	0.5008	0.083	0.735
inc. Gene-Treat. Interaction**	112.2574	0.2488	0.0208	0.000	1.8768	0.6478	0.322	0.0053	0.5834	0.739	0.082
(d) <i>Binary Phenotype</i> ***	-2.5510	0.0365	0.0031	0.000	0.2715	0.0958	0	0.0002	0.0739	0.791	0.071
Fixed Treat Effect (c = 5)	111.7150	0.3382	0.0224	0.064	2.5144	0.7096	0.688	1.3056	0.5474	0.142	0.664
(e) Fixed Treat Effect (c = 10)	111.4419	0.3689	0.0238	0.483	2.7390	0.7552	0.771	1.3065	0.5826	0.175	0.611
Fixed Treat Effect (c = 15)	111.1688	0.3996	0.0255	0.800	2.9636	0.8104	0.79	1.3075	0.6251	0.214	0.546
(f) Fixed Substitution (m=130)	115.0741	0.2242	0.0187	0.000	1.6639	0.5937	0.176	0.0005	0.4580	0.815	0.039
Fixed Substitution (m=140)	114.5279	0.2856	0.0200	0.000	2.1130	0.6358	0.442	0.0023	0.4904	0.810	0.052
(g) Random Substitution	113.9789	0.3470	0.0233	0.168	2.5648	0.7401	0.719	0.0065	0.5709	0.786	0.052
Median Method (k = 160)	106.7096	0.5177	0.0633	0.295	4.0900	1.4484	0.688	0.0374	1.0882	0.788	0.023
(h) Median Method (k = 180)	106.7058	0.5178	0.0672	0.304	4.0942	1.4720	0.689	0.0361	1.1115	0.796	0.018
Median Method (k = 200)	106.7050	0.5178	0.0671	0.304	4.0926	1.4699	0.693	0.0370	1.1080	0.796	0.022
(i) Non-parametric Adjustment	111.6022	0.4097	0.0261	0.757	3.0342	0.8294	0.78	1.0247	0.6398	0.377	0.359
(j) Censored Normal Regression	109.6945	0.4300	0.0278	0.572	3.1997	0.8744	0.757	0.9335	0.6728	0.462	0.291

Table 43: Results for 1,000 runs of Scenario 5a when $\beta_3 = 0$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 14.2627, with a mean standard error of 0.7542.

**Approach C, when a gene-treatment interaction is modelled, yields a mean treatment coefficient of 10.4963 with mean standard error of 1.2818; and mean gene-treatment interaction of 3.7649 with mean standard error of 1.0377.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	110.0279	0.3991	0.0253	0.800	2.9904	0.8053	0.802	2.0085	0.6216	0.826	0.917
(a) No Adjustment	111.9443	0.3082	0.0214	0.001	2.3297	0.6784	0.598	0.2563	0.5236	0.021	0.068
(b) Exclude	109.4724	0.2919	0.0260	0.003	2.2222	0.8175	0.619	1.4945	0.6364	0.672	0.654
(c) Treatment as binary covariate*	112.7115	0.2485	0.0210	0.000	1.8954	0.6512	0.329	-0.0353	0.5025	0.005	0.042
inc. Gene-Treat. Interaction**	111.8149	0.2491	0.0209	0.000	1.8972	0.6483	0.329	1.4935	0.5933	0.630	0.709
(d) <i>Binary Phenotype</i> ***	-2.7298	0.0363	0.0032	.	0.2727	0.0967	.	0.1846	0.0741	.	0.718
Fixed Treat Effect (c = 5)	111.5545	0.3384	0.0224	0.072	2.5501	0.7125	0.715	0.4042	0.5499	0.047	0.110
(e) Fixed Treat Effect (c = 10)	111.1647	0.3685	0.0239	0.488	2.7705	0.7577	0.782	0.5522	0.5849	0.106	0.143
Fixed Treat Effect (c = 15)	110.7749	0.3987	0.0256	0.806	2.9908	0.8123	0.798	0.7001	0.6270	0.190	0.173
(f) Fixed Substitution (m=130)	113.6967	0.2280	0.0188	0.000	1.7170	0.5978	0.195	1.1397	0.4615	0.268	0.692
Fixed Substitution (m=140)	112.9172	0.2883	0.0202	0.000	2.1578	0.6414	0.489	1.4355	0.4951	0.536	0.838
(g) Random Substitution	112.1218	0.3489	0.0234	0.173	2.6041	0.7448	0.714	1.7344	0.5749	0.755	0.861
Median Method (k = 160)	105.8057	0.4834	0.0581	0.457	3.8695	1.3671	0.724	2.6461	1.0762	0.725	0.547
(h) Median Method (k = 180)	105.7864	0.4837	0.0604	0.476	3.8745	1.3831	0.721	2.6532	1.1066	0.730	0.514
Median Method (k = 200)	105.7868	0.4837	0.0603	0.479	3.8753	1.3810	0.718	2.6536	1.1069	0.734	0.524
(i) Non-parametric Adjustment	110.8284	0.4069	0.0260	0.769	3.0438	0.8275	0.794	1.0165	0.6387	0.383	0.351
(j) Censored Normal Regression	108.9111	0.4253	0.0275	0.620	3.1967	0.8684	0.781	1.2279	0.6741	0.537	0.432

Table 44: Results for 1,000 runs of Scenario 5b when $\beta_3 = 2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 1.9499, with a mean standard error of 0.7552.

**Approach C, when a gene-treatment interaction is additionally modelled, yields a mean treatment coefficient of 10.4719 with mean standard error of 1.2941; and mean gene-treatment interaction of -8.2658 with mean standard error of 1.0273.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	112.3793	0.4004	0.0254	0.795	2.9634	0.8052	0.801	-1.9883	0.6213	0.817	0.900
(a) No Adjustment	113.8109	0.3099	0.0214	0.001	2.3221	0.6783	0.605	-2.7823	0.5234	0.403	1.000
(b) Exclude	111.2185	0.2930	0.0261	0.000	2.2287	0.8179	0.633	-1.4629	0.6259	0.637	0.629
(c) Treatment as binary covariate*	114.2525	0.2485	0.0209	0.000	1.8851	0.6497	0.336	-2.4833	0.5011	0.599	0.998
inc. Gene-Treat. Interaction**	111.2185	0.2930	0.0261	0.000	2.2287	0.8179	0.633	-1.4629	0.6259	0.637	0.629
(d) <i>Binary Phenotype</i> ***	-2.5199	0.0366	0.0032	.	0.2718	0.0967	.	-0.1830	0.0753	.	0.667
Fixed Treat Effect (c = 5)	113.5915	0.3402	0.0224	0.082	2.5378	0.7131	0.691	-2.9295	0.5502	0.328	1.000
(e) Fixed Treat Effect (c = 10)	113.3722	0.3705	0.0239	0.513	2.7535	0.7589	0.784	-3.0766	0.5855	0.275	1.000
Fixed Treat Effect (c = 15)	113.1528	0.4008	0.0256	0.787	2.9692	0.8139	0.81	-3.2238	0.6280	0.235	1.000
(f) Fixed Substitution (m=130)	115.0384	0.2289	0.0188	0.000	1.7167	0.5983	0.187	-1.1451	0.4616	0.292	0.698
Fixed Substitution (m=140)	114.5996	0.2895	0.0202	0.000	2.1480	0.6417	0.485	-1.4394	0.4951	0.562	0.833
(g) Random Substitution	114.1583	0.3502	0.0235	0.188	2.5794	0.7452	0.739	-1.7350	0.5749	0.774	0.861
Median Method (k = 160)	108.7401	0.4877	0.0592	0.455	3.8452	1.3733	0.739	-2.5299	1.0210	0.745	0.566
(h) Median Method (k = 180)	108.7317	0.4879	0.0610	0.458	3.8484	1.3821	0.742	-2.5321	1.0265	0.750	0.568
Median Method (k = 200)	108.7322	0.4879	0.0607	0.453	3.8494	1.3810	0.731	-2.5316	1.0257	0.753	0.566
(i) Non-parametric Adjustment	113.2257	0.4110	0.0262	0.735	3.0337	0.8330	0.802	-3.0084	0.6427	0.381	0.998
(j) Censored Normal Regression	111.4479	0.4273	0.0276	0.595	3.1761	0.8701	0.791	-2.9914	0.6699	0.407	0.993

Table 45: Results for 1,000 runs of Scenario 5b, when $\beta_3 = -2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 4.4101, with a mean standard error of 0.8040.

**The additional analysis for (c) which includes the gene-treatment interaction term yields a mean treatment coefficient of 10.4916 with mean standard error of 1.3241; and mean gene-treatment interaction of -6.7010 with mean standard error of 1.1042.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Type I Error
Underlying SBP	112.0534	0.3987	0.0253	0.813	3.0425	0.8054	0.799	0.0133	0.6216	0.810	0.047
(a) No Adjustment	113.6585	0.3067	0.0213	0.000	2.3559	0.6774	0.612	-1.3231	0.5229	0.107	0.719
(b) Exclude	110.8960	0.2931	0.0263	0.002	2.2740	0.8267	0.641	0.0086	0.6376	0.831	0.045
(c) Treatment as binary covariate*	114.1789	0.2473	0.0209	0.000	1.9113	0.6506	0.366	-1.3261	0.5015	0.078	0.756
inc. Gene-Treat. Interaction**	112.2425	0.2494	0.0209	0.000	1.8729	0.6484	0.33	0.0023	0.5846	0.784	0.075
(d) <i>Binary Phenotype</i> ***	-2.5436	0.0364	0.0031	.	0.2790	0.0958	.	-0.0006	0.0739	.	0.048
Fixed Treat Effect (c = 5)	113.3884	0.3374	0.0224	0.065	2.5858	0.7119	0.733	-1.3211	0.5495	0.124	0.688
(e) Fixed Treat Effect (c = 10)	113.1182	0.3681	0.0239	0.483	2.8157	0.7579	0.798	-1.3191	0.5850	0.153	0.620
Fixed Treat Effect (c = 15)	112.8480	0.3987	0.0256	0.807	3.0456	0.8134	0.791	-1.3171	0.6278	0.200	0.554
(f) Fixed Substitution (m=130)	115.1668	0.2223	0.0187	0.000	1.7065	0.5939	0.194	0.0097	0.4584	0.833	0.044
Fixed Substitution (m=140)	114.6265	0.2836	0.0200	0.000	2.1663	0.6358	0.483	0.0137	0.4907	0.827	0.053
(g) Random Substitution	114.0910	0.3448	0.0233	0.135	2.6309	0.7402	0.73	0.0176	0.5713	0.823	0.046
Median Method (k = 160)	106.7914	0.5165	0.0635	0.299	4.1793	1.4381	0.653	0.0316	1.0838	0.804	0.023
(h) Median Method (k = 180)	106.7702	0.5169	0.0674	0.302	4.1857	1.4683	0.663	0.0315	1.1048	0.806	0.020
Median Method (k = 200)	106.7715	0.5169	0.0675	0.302	4.1836	1.4654	0.663	0.0308	1.1028	0.803	0.021
(i) Non-parametric Adjustment	112.9358	0.4094	0.0262	0.771	3.1216	0.8337	0.789	-1.0152	0.6435	0.380	0.348
(j) Censored Normal Regression	110.9175	0.4292	0.0278	0.567	3.2879	0.8777	0.769	-0.9207	0.6784	0.463	0.261

Table 46: Results for 1,000 runs of Scenario 5b, when $\beta_3 = 0$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 3.0115, with a mean standard error of 0.7758.

**Approach C, when a gene-treatment interaction is modelled, yields a mean treatment coefficient of 10.5020 with mean standard error of 1.2070; and mean gene-treatment interaction of -7.4913 with mean standard error of 1.0587.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	109.9856	0.3998	0.0254	0.814	3.0144	0.8059	0.791	2.0451	0.6220	0.788	0.905
(a) No Adjustment	110.8905	0.3093	0.0213	0.000	2.3489	0.6764	0.6	2.0284	0.5221	0.810	0.972
(b) Exclude	109.4190	0.2931	0.0261	0.002	2.2468	0.8183	0.615	1.5167	0.6368	0.685	0.660
(c) Treatment as binary covariate* inc. Gene-Treat. Interaction**	111.6569	0.2494	0.0209	0.000	1.9083	0.6486	0.362	1.7252	0.5005	0.729	0.925
(d) <i>Binary Phenotype</i> ***	-2.7321	0.0364	0.0032	.	0.2732	0.0966	.	0.1873	0.0741	.	0.722
Fixed Treat Effect (c = 5)	110.5047	0.3394	0.0224	0.059	2.5700	0.7108	0.701	2.1801	0.5487	0.784	0.978
(e) Fixed Treat Effect (c = 10)	110.1189	0.3694	0.0238	0.503	2.7911	0.7565	0.776	2.3318	0.5839	0.735	0.980
Fixed Treat Effect (c = 15)	109.7331	0.3995	0.0255	0.807	3.0121	0.8114	0.796	2.4835	0.6263	0.680	0.981
(f) Fixed Substitution (m=130)	113.6568	0.2288	0.0188	0.000	1.7404	0.5982	0.206	1.1552	0.4617	0.288	0.705
Fixed Substitution (m=140)	112.8852	0.2890	0.0202	0.000	2.1825	0.6418	0.496	1.4586	0.4954	0.567	0.827
(g) Random Substitution	112.1263	0.3488	0.0235	0.155	2.6320	0.7453	0.731	1.7606	0.5753	0.743	0.862
Median Method (k = 160)	105.7247	0.4849	0.0582	0.444	3.8833	1.3708	0.716	2.7231	1.0828	0.722	0.582
(h) Median Method (k = 180)	105.7036	0.4853	0.0607	0.461	3.8869	1.3859	0.722	2.7299	1.1133	0.730	0.571
Median Method (k = 200)	105.7039	0.4852	0.0606	0.467	3.8856	1.3928	0.723	2.7310	1.1151	0.726	0.559
(i) Non-parametric Adjustment	109.9893	0.4083	0.0261	0.779	3.0722	0.8284	0.79	2.4297	0.6394	0.706	0.972
(j) Censored Normal Regression	108.1776	0.4255	0.0275	0.641	3.2164	0.8677	0.785	2.5065	0.6715	0.686	0.965

Table 47: Results for 1,000 runs of Scenario 6a, when $\beta_3 = 2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 10.9792, with a mean standard error of 0.7331.

**Approach C, when a gene-treatment interaction is additionally modelled, yields a mean treatment coefficient of 10.4981 with mean standard error of 1.2765; and mean gene-treatment interaction of 0.4665 with mean standard error of 1.0124.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	112.3728	0.3996	0.0254	0.814	2.9995	0.8067	0.803	-1.9818	0.6228	0.806	0.889
(a) No Adjustment	112.7608	0.3089	0.0213	0.001	2.3343	0.6776	0.589	-1.1489	0.5232	0.369	0.580
(b) Exclude	111.1979	0.2928	0.0261	0.004	2.2305	0.8189	0.609	-1.4739	0.6270	0.655	0.648
(c) Treatment as binary covariate*	113.1840	0.2492	0.0209	0.000	1.8918	0.6503	0.331	-0.8640	0.5019	0.162	0.402
inc. Gene-Treat. Interaction**	113.5582	0.2492	0.0209	0.000	1.8899	0.6498	0.327	-1.4664	0.5849	0.609	0.686
(d) <i>Binary Phenotype</i> ***	-2.5142	0.0364	0.0032	.	0.2728	0.0967	.	-0.1808	0.0753	.	0.664
Fixed Treat Effect (c = 5)	112.5468	0.3389	0.0224	0.069	2.5569	0.7118	0.713	-1.2928	0.5496	0.486	0.638
(e) Fixed Treat Effect (c = 10)	112.3328	0.3690	0.0238	0.494	2.7795	0.7572	0.785	-1.4367	0.5846	0.607	0.680
Fixed Treat Effect (c = 15)	112.1188	0.3991	0.0256	0.806	3.0022	0.8119	0.815	-1.5806	0.6268	0.693	0.707
(f) Fixed Substitution (m=130)	115.0187	0.2289	0.0189	0.000	1.7248	0.5993	0.205	-1.1506	0.4627	0.306	0.687
Fixed Substitution (m=140)	114.5907	0.2890	0.0202	0.000	2.1700	0.6429	0.489	-1.4384	0.4964	0.552	0.818
(g) Random Substitution	114.1698	0.3491	0.0235	0.186	2.6105	0.7462	0.744	-1.7257	0.5761	0.746	0.844
Median Method (k = 160)	108.7664	0.4853	0.0592	0.439	3.9025	1.3758	0.693	-2.4886	1.0216	0.754	0.557
(h) Median Method (k = 180)	108.7647	0.4854	0.0605	0.439	3.9028	1.3839	0.707	-2.4901	1.0288	0.759	0.559
Median Method (k = 200)	108.7630	0.4854	0.0606	0.443	3.9034	1.3889	0.709	-2.4899	1.0268	0.759	0.558
(i) Non-parametric Adjustment	112.4236	0.4069	0.0260	0.786	3.0553	0.8266	0.815	-1.6990	0.6382	0.748	0.758
(j) Censored Normal Regression	110.7164	0.4253	0.0275	0.623	3.2106	0.8679	0.782	-1.8210	0.6675	0.785	0.773

Table 48: Results for 1,000 runs of Scenario 6a, when $\beta_3 = -2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 12.4370, with a mean standard error of 0.7897.

**Approach C, when a gene-treatment interaction is additionally modelled, yields a mean treatment coefficient of 10.4751 with mean standard error of 1.3100; and mean gene-treatment interaction of 2.0499 with mean standard error of 1.0920.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Type I Error
Underlying SBP	111.2457	0.3995	0.0254	0.824	2.9796	0.8061	0.816	0.0000	0.6218	0.793	0.050
(a) No Adjustment	111.8821	0.3089	0.0213	0.000	2.2996	0.6763	0.566	0.4257	0.5217	0.663	0.130
(b) Exclude	110.3291	0.2934	0.0260	0.001	2.1764	0.8172	0.59	0.0100	0.6301	0.794	0.040
(c) Treatment as binary covariate*	112.4729	0.2490	0.0209	0.000	1.8496	0.6485	0.296	0.4282	0.4996	0.640	0.140
inc. Gene-Treat. Interaction**	112.7220	0.2490	0.0209	0.000	1.8496	0.6484	0.295	0.0083	0.5878	0.767	0.059
(d) <i>Binary Phenotype</i> ***	-2.6183	0.0363	0.0032	.	0.2764	0.0965	.	0.0003	0.0744	.	0.042
Fixed Treat Effect (c = 5)	111.5849	0.3390	0.0224	0.068	2.5253	0.7108	0.703	0.4242	0.5483	0.671	0.126
(e) Fixed Treat Effect (c = 10)	111.2877	0.3690	0.0238	0.488	2.7510	0.7565	0.783	0.4227	0.5836	0.691	0.111
Fixed Treat Effect (c = 15)	110.9905	0.3991	0.0255	0.827	2.9767	0.8114	0.824	0.4212	0.6259	0.705	0.105
(f) Fixed Substitution (m=130)	114.3493	0.2294	0.0188	0.000	1.6878	0.5978	0.181	0.0055	0.4612	0.801	0.044
Fixed Substitution (m=140)	113.7549	0.2895	0.0202	0.000	2.1392	0.6417	0.462	0.0025	0.4951	0.806	0.047
(g) Random Substitution	113.1663	0.3495	0.0235	0.181	2.5913	0.7454	0.736	-0.0007	0.5750	0.792	0.050
Median Method (k = 160)	107.4695	0.4811	0.0596	0.484	3.8979	1.3709	0.739	-0.0189	1.0326	0.798	0.022
(h) Median Method (k = 180)	107.4709	0.4811	0.0605	0.487	3.8953	1.3755	0.745	-0.0187	1.0381	0.800	0.020
Median Method (k = 200)	107.4691	0.4811	0.0606	0.486	3.8968	1.3752	0.746	-0.0192	1.0410	0.799	0.019
(i) Non-parametric Adjustment	111.2732	0.4071	0.0260	0.771	3.0315	0.8265	0.817	0.3319	0.6376	0.729	0.083
(j) Censored Normal Regression	109.5156	0.4251	0.0275	0.647	3.1799	0.8668	0.808	0.3037	0.6680	0.744	0.074

Table 49: Results for 1,000 runs of Scenario 6a, when $\beta_3 = 0$.

Mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and type I error relative to the genetic factor (at 5% level of sig.) are shown.

*Approach C also yields a mean treatment coefficient of 11.7474, with a mean standard error of 0.7560.

**Approach C, when a gene-treatment interaction is additionally modelled, yields a mean treatment coefficient of 10.4527 with mean standard error of 1.2895; and mean gene-treatment interaction of 1.2935 with mean standard error of 1.0443.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	110.0888	0.3988	0.0254	0.795	2.9890	0.8054	0.802	2.0386	0.6214	0.772	0.893
(a) No Adjustment	111.4879	0.3085	0.0213	0.002	2.3330	0.6768	0.61	1.1407	0.5222	0.380	0.577
(b) Exclude	109.4781	0.2929	0.0261	0.001	2.2360	0.8187	0.625	1.4941	0.6373	0.648	0.647
(c) Treatment as binary covariate* inc. Gene-Treat. Interaction**	112.2496	0.2487	0.0209	0.000	1.8970	0.6492	0.327	0.8347	0.5008	0.158	0.391
(d) <i>Binary Phenotype</i> ***	-2.7240	0.0364	0.0032	.	0.2718	0.0966	.	0.1875	0.0741	.	0.719
Fixed Treat Effect (c = 5)	111.1037	0.3385	0.0224	0.072	2.5527	0.7112	0.725	1.2949	0.5488	0.492	0.652
(e) Fixed Treat Effect (c = 10)	110.7195	0.3686	0.0238	0.491	2.7725	0.7567	0.782	1.4490	0.5839	0.621	0.684
Fixed Treat Effect (c = 15)	110.3353	0.3987	0.0256	0.795	2.9922	0.8116	0.809	1.6031	0.6262	0.693	0.716
(f) Fixed Substitution (m=130)	113.7182	0.2284	0.0188	0.000	1.7209	0.5981	0.189	1.1396	0.4615	0.287	0.692
Fixed Substitution (m=140)	112.9498	0.2885	0.0202	0.000	2.1603	0.6417	0.478	1.4479	0.4951	0.546	0.822
(g) Random Substitution	112.1832	0.3487	0.0235	0.180	2.6008	0.7452	0.747	1.7548	0.5750	0.731	0.856
Median Method (k = 160)	105.7130	0.4864	0.0583	0.429	3.9317	1.3678	0.718	2.6935	1.0789	0.718	0.579
(h) Median Method (k = 180)	105.6887	0.4868	0.0610	0.441	3.9377	1.3885	0.726	2.7008	1.1140	0.734	0.541
Median Method (k = 200)	105.6885	0.4868	0.0609	0.437	3.9371	1.3942	0.731	2.7023	1.1128	0.733	0.558
(i) Non-parametric Adjustment	110.4914	0.4068	0.0260	0.768	3.0476	0.8271	0.8	1.7330	0.6382	0.740	0.765
(j) Censored Normal Regression	108.6175	0.4251	0.0275	0.624	3.2008	0.8681	0.799	1.8785	0.6728	0.774	0.775

Table 50: Results for 1,000 runs of Scenario 6b, when $\beta_3 = 2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 7.1338, with a mean standard error of 0.7475.

**Approach C, when a gene-treatment interaction is additionally modelled, yields a mean treatment coefficient of 10.5041 with mean standard error of 1.2974; and mean gene-treatment interaction of -3.2705 with mean standard error of 1.0300.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Power
Underlying SBP	112.5000	0.3985	0.0254	0.800	3.0021	0.8057	0.79	-1.9998	0.6217	0.786	0.888
(a) No Adjustment	113.3853	0.3085	0.0213	0.002	2.3347	0.6771	0.6	-1.9777	0.5225	0.781	0.966
(b) Exclude	111.2775	0.2925	0.0261	0.001	2.2261	0.8186	0.603	-1.4880	0.6267	0.668	0.664
(c) Treatment as binary covariate*	113.8008	0.2481	0.0209	0.000	1.8900	0.6490	0.328	-1.6861	0.5006	0.697	0.919
inc. Gene-Treat. Interaction**	113.6677	0.2483	0.0209	0.000	1.8903	0.6491	0.334	-1.4801	0.5845	0.624	0.707
(d) <i>Binary Phenotype</i> ***	-2.5083	0.0364	0.0032	.	0.2739	0.0966	.	-0.1829	0.0753	.	0.688
Fixed Treat Effect (c = 5)	113.1774	0.3385	0.0224	0.067	2.5561	0.7117	0.709	-2.1226	0.5492	0.781	0.972
(e) Fixed Treat Effect (c = 10)	112.9695	0.3686	0.0239	0.475	2.7776	0.7575	0.759	-2.2675	0.5845	0.757	0.970
Fixed Treat Effect (c = 15)	112.7615	0.3987	0.0256	0.805	2.9990	0.8125	0.785	-2.4124	0.6269	0.713	0.965
(f) Fixed Substitution (m=130)	115.1055	0.2281	0.0188	0.000	1.7226	0.5984	0.204	-1.1636	0.4617	0.298	0.712
Fixed Substitution (m=140)	114.6896	0.2882	0.0202	0.000	2.1655	0.6419	0.485	-1.4534	0.4953	0.559	0.830
(g) Random Substitution	114.2685	0.3483	0.0235	0.181	2.6133	0.7452	0.73	-1.7380	0.5750	0.742	0.855
Median Method (k = 160)	108.9086	0.4842	0.0591	0.464	3.8712	1.3721	0.726	-2.4885	1.0208	0.756	0.548
(h) Median Method (k = 180)	108.9060	0.4843	0.0606	0.472	3.8723	1.3779	0.73	-2.4902	1.0243	0.757	0.550
Median Method (k = 200)	108.9067	0.4843	0.0606	0.471	3.8720	1.3839	0.73	-2.4911	1.0261	0.759	0.547
(i) Non-parametric Adjustment	112.9524	0.4073	0.0261	0.775	3.0588	0.8290	0.786	-2.3619	0.6397	0.731	0.953
(j) Censored Normal Regression	111.2112	0.4249	0.0276	0.634	3.2029	0.8687	0.765	-2.4194	0.6684	0.713	0.948

Table 51: Results for 1,000 runs of Scenario 6b, when $\beta_3 = -2$.

Shown are mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power relative to the genetic factor (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 8.8731, with a mean standard error of 0.7985.

**Approach C, when a gene-treatment interaction is additionally modelled, yields a mean treatment coefficient of 10.5177 with mean standard error of 1.3267; and mean gene-treatment interaction of -1.7162 with mean standard error of 1.1057.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Gene	Gene SE	Gene Coverage	Type I Error
Underlying SBP	111.2692	0.3983	0.0254	0.779	3.0198	0.8059	0.811	0.0106	0.6216	0.795	0.048
(a) No Adjustment	112.4307	0.3079	0.0213	0.001	2.3528	0.6771	0.617	-0.4143	0.5223	0.669	0.122
(b) Exclude	110.3998	0.2914	0.0261	0.001	2.2420	0.8184	0.626	0.0199	0.6312	0.806	0.051
(c) Treatment as binary covariate* inc. Gene-Treat. Interaction**	113.0436	0.2474	0.0209	0.000	1.9024	0.6490	0.331	-0.4143	0.5000	0.666	0.129
(d) <i>Binary Phenotype</i> ***	-2.6175	0.0364	0.0032	.	0.2753	0.0965	.	-0.0003	0.0744	.	0.053
Fixed Treat Effect (c = 5)	112.1252	0.3381	0.0224	0.085	2.5772	0.7117	0.738	-0.4142	0.5490	0.677	0.109
(e) Fixed Treat Effect (c = 10)	111.8197	0.3683	0.0239	0.474	2.8016	0.7575	0.802	-0.4141	0.5843	0.691	0.109
Fixed Treat Effect (c = 15)	111.5142	0.3985	0.0256	0.783	3.0259	0.8125	0.81	-0.4140	0.6267	0.703	0.095
(f) Fixed Substitution (m=130)	114.4100	0.2278	0.0188	0.000	1.7327	0.5983	0.205	0.0173	0.4615	0.807	0.052
Fixed Substitution (m=140)	113.7990	0.2882	0.0202	0.000	2.1815	0.6422	0.494	0.0176	0.4954	0.803	0.046
(g) Random Substitution	113.1865	0.3485	0.0235	0.187	2.6343	0.7459	0.748	0.0214	0.5754	0.803	0.051
Median Method (k = 160)	107.4135	0.4819	0.0596	0.478	3.8961	1.3720	0.722	0.0327	1.0440	0.823	0.021
(h) Median Method (k = 180)	107.4128	0.4819	0.0606	0.483	3.8959	1.3778	0.722	0.0322	1.0476	0.817	0.022
Median Method (k = 200)	107.4133	0.4819	0.0607	0.491	3.8948	1.3798	0.724	0.0317	1.0451	0.821	0.020
(i) Non-parametric Adjustment	111.6799	0.4066	0.0261	0.745	3.0801	0.8278	0.816	-0.3201	0.6385	0.739	0.075
(j) Censored Normal Regression	109.8870	0.4245	0.0276	0.636	3.2327	0.8684	0.793	-0.2872	0.6701	0.764	0.069

Table 52: Results for 1,000 runs of Scenario 6b, when $\beta_3 = 0$.

Mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and type I error relative to the genetic factor (at 5% level of sig.) are shown.

*Approach C also yields a mean treatment coefficient of 8.0420, with a mean standard error of 0.7683.

**Approach C, when a gene-treatment interaction is additionally modelled, yields a mean treatment coefficient of 10.5314 with mean standard error of 1.3087; and mean gene-treatment interaction of -2.4914 with mean standard error of 1.0607.

*** Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Analysis	Int	Age	Age SE	Age Coverage	Sex	Sex SE	Sex Coverage	Diabetes	Diabetes SE	Gene	Gene SE	Gene Coverage	Gene2	Gene2 SE	Gene2 Coverage	Power	Type I Error
Underlying SBP	110.0215	0.3992	0.0254	0.782	3.0652	0.8060	0.792	-0.0417	1.0075	2.0011	0.6220	0.819	-0.0003	0.6222	0.807	0.904	0.055
(a) No Adjustment	111.5642	0.3028	0.0209	0.000	2.3365	0.6655	0.586	-4.0424	0.8319	1.5185	0.5136	0.623	-0.0044	0.5138	0.801	0.850	0.047
(b) Exclude	110.3096	0.2778	0.0265	0.000	2.1466	0.8292	0.589	-7.8638	1.2109	1.3953	0.6462	0.626	0.0159	0.6395	0.796	0.570	0.046
(c) Treatment as binary covariate*	112.6637	0.2341	0.0204	0.000	1.8196	0.6316	0.274	-6.8915	0.8108	1.1762	0.4872	0.343	-0.0085	0.4869	0.803	0.674	0.043
(d) Binary Phenotype **	-2.7255	0.0362	0.0031	.	0.2774	0.0962	.	0.8124	0.1188	0.1817	0.0739	.	0.0037	0.0742	.	0.700	0.057
Fixed Treat Effect (c = 5)	111.0485	0.3349	0.0221	0.054	2.5788	0.7039	0.708	-2.7073	0.8799	1.6790	0.5432	0.731	-0.0025	0.5434	0.808	0.882	0.051
(e) Fixed Treat Effect (c = 10)	110.5329	0.3671	0.0237	0.457	2.8210	0.7536	0.776	-1.3722	0.9420	1.8394	0.5815	0.792	-0.0007	0.5817	0.813	0.893	0.056
Fixed Treat Effect (c = 15)	110.0172	0.3993	0.0256	0.778	3.0632	0.8126	0.792	-0.0371	1.0157	1.9999	0.6271	0.812	0.0012	0.6273	0.810	0.893	0.058
(f) Fixed Substitution (m=130)	114.3115	0.2179	0.0180	0.000	1.6730	0.5722	0.153	-2.8843	0.7153	1.0735	0.4416	0.192	0.0116	0.4417	0.797	0.683	0.047
Fixed Substitution (m=140)	113.2802	0.2822	0.0199	0.000	2.1574	0.6334	0.478	-0.2142	0.7917	1.3944	0.4888	0.507	0.0154	0.4889	0.801	0.824	0.051
(g) Random Substitution	112.2401	0.3468	0.0237	0.177	2.6386	0.7549	0.73	2.4463	0.9436	1.7142	0.5826	0.763	0.0216	0.5828	0.806	0.856	0.052
Median Method (k = 160)	103.4481	0.5364	0.0686	0.273	3.8617	1.4385	0.726	11.5652	2.9848	2.5179	1.0748	0.762	0.0387	1.0460	0.813	0.522	0.016
(h) Median Method (k = 180)	101.9367	0.5644	0.0901	0.291	4.1443	1.6394	0.727	29.5733	5.7008	2.7130	1.2463	0.767	0.0271	1.1595	0.831	0.435	0.011
Median Method (k = 200)	101.7877	0.5672	0.0931	0.293	4.1684	1.6669	0.729	48.6045	8.9549	2.7317	1.2722	0.769	0.0266	1.1765	0.824	0.395	0.010
(i) Non-parametric Adjustment	110.0851	0.4158	0.0267	0.694	3.1803	0.8474	0.783	1.8686	1.0593	2.0766	0.6540	0.807	0.0081	0.6542	0.807	0.898	0.057
(j) Censored Normal Regression	107.5264	0.4417	0.0287	0.420	3.3935	0.9040	0.746	1.6481	1.1991	2.2255	0.7006	0.781	0.0115	0.6976	0.805	0.894	0.059

Table 53: Results for 1,000 runs of Scenario 7, when the probability of treatment is differential by exposure to diabetes.

Mean parameter estimates/SEs, coverage rates based on 80% confidence intervals, and power and type I error relative to g1 and g2 respectively (at 5% level of sig.).

*Approach C also yields a mean treatment coefficient of 10.3917, with a mean standard error of 0.7021.

**Parameter estimates for Approach D are log odds-ratios, and are non-comparable to the other approaches.

Appendix B.

B.1. NIH Background Fact Sheet on GWAS Policy

Update

August 28, 2008

A research team, led by David W. Craig, Ph.D. at the Translational Genomics Research Institute (TGen) in Phoenix AZ, has developed a new bioinformatics method that allows the detection of a single person's SNP profile in a mixture of 1,000 or more individual DNA samples. In other words, bioinformatics techniques have progressed to the point that with enough genomic data on an individual from another source, it is now possible to determine whether that individual participated in a study by analyzing only the pooled summary data.

SNP stands for single nucleotide polymorphism, which is a change in a genetic letter in a specific location on a DNA molecule when compared to other DNA molecules. SNPs are used to study human genetic variation and are a powerful way to investigate genetic predispositions to health or disease. Large-scale genomic studies of human variation – called genome-wide association studies or GWAS – have recently provided important clues to the genetic roots of numerous common diseases. Because of the power of this technology, many institutes and centers at the National Institutes of Health support or are involved in such studies to understand the genetics of common maladies.

This new bioinformatics method is powerful, but it is still not simple to detect a specific individual's SNP profile in a pooled dataset. To find a specific profile within a set, the inquirer would first need to already have a highly-dense genomic profile (currently at least 10,000 SNPs) from an individual. Then this SNP profile would need to be statistically compared against the study dataset to measure how similar or different it is. Prior expectations were that individual profiles would have to be compared one to one to confirm a match; however, this new statistical analysis can now be used to detect a profile even in pooled data.

Although the technique has been demonstrated to work, the NIH is unaware that it has been used to compromise any information within NIH GWAS datasets. The technology to obtain the required genomic profile is not commonly used outside of the research community. And, even if an individual's SNP profile was found within a pooled dataset, all that would be learned is that this profile was contained in the dataset and, thus, it could then be associated with the characteristics of that dataset (e.g., disease or control population). The NIH GWAS databases do not contain the names or other identifiable information about individual study participants, so there is no risk to an individual participant's financial accounts or other personal information.

This discovery, however, has important policy implications for the way the scientific community shares such pooled sets of genetic data. For example, scientific journals have required researchers to make available aggregate data from GWAS studies when the results are published as a means to ensure the quality of the data. And, because use of these pooled datasets can speed up disease gene discovery, NIH – as well as other research institutions and individual laboratories – developed public databases that allow researchers to freely download the datasets into their computers for analysis.

Because individual SNP profiles can now be detected within aggregate data, the NIH has moved quickly to assure continued protection of research participant privacy in genomics studies by controlling access to pooled datasets. For example, on Monday, Aug. 25, 2008, the NIH removed aggregate statistics files of individual GWAS studies from the public portion of the databases it manages, such as the Database of Genotypes and Phenotypes (dbGaP), operated by the National Center for Biotechnology Information, and the Cancer Genetic Markers of Susceptibility (CGEMS), operated by the National Cancer Institute. The data remains available for researcher use, but researchers must now apply for access to the data and agree to protect the confidentiality of the data in the same way that has been done all along for individual-level study data.

In addition, NIH is aware that others operating databases with similar types of datasets, including the Wellcome Trust Case Control Consortium in England and the Broad Institute of MIT and Harvard in Boston, have removed aggregate data from public availability.

NIH will continue to focus on this fast-moving field of research and on the development of policies to appropriately manage its databases and to promote policies that protect the confidentiality of all those who participate in NIH-sponsored research studies.

B.2. Breakdown and Proof of the Visscher *et al.* Linear Regression Approach

The following notation, which is consistent with that outlined in Section 2.5, is to be used throughout. The j 'th SNP ($j = 1, \dots, s$) has population allele frequency p_j ; the allele frequency in the mixture is denoted q_j , and the allele frequency in the reference group is denoted r_j . The mixture consists of N_{mix} individuals, and the reference group consists of N_{ref} individuals. An estimate of p_j – based on the combined samples of N_{mix} and N_{ref} individuals – is denoted \hat{p}_j . The individual of interest – individual i – has scaled genotype y_{ij} ($= 0, 0.5$ or 1).

B.2.1 Population frequencies known:

A regression of Y_{ij} on X_j is fitted, where $Y_{ij} = y_{ij} - p_j$, and $X_j = q_j - p_j$.

Visscher *et al.* state that the regression coefficient b_j is estimated by

$$\hat{b}_j = [X'X]^{-1}X'y = \frac{\sum_{j=1}^s (y_{ij} - p_j)(q_j - p_j)}{\sum_{j=1}^s (q_j - p_j)^2}.$$

This is the least squares estimator for a general linear model with no intercept (i.e. where the design matrix, X , consists only of the explanatory variable X_j). Note, however, that including an intercept term in the model will have no tangible effect on b , and in my practical illustrations of the method, i.e. in sections 2.6 and 2.7, an intercept term *is* fitted.

The genotype, y_{ij} , can be expressed as $\frac{1}{2}g_{ij}$, where g_{ij} is the sum of the two alleles a_{1ij} and a_{2ij} ($= 0$ or 1 copy of the minor allele). a_{1ij} and a_{2ij} are both

Bernoulli distributed with probability p_j and variance $p_j(1-p_j)$ and, thus, Y_{ij} has variance:

$$\begin{aligned} \text{Var}(Y_{ij}) &= \text{Var}(y_{ij} - p_j) = \text{Var}\left(\frac{1}{2}g_{ij}\right) = \left(\frac{1}{2}\right)^2 \text{Var}(g_{ij}) \\ &= \left(\frac{1}{2}\right)^2 [\text{Var}(a_{1ij}) + \text{Var}(a_{2ij})] = \left(\frac{1}{2}\right)^2 \cdot 2 \cdot p_j(1 - p_j) \\ &= \frac{p_j(1-p_j)}{2} \end{aligned}$$

Similarly, q_j can be expressed as $\sum_{i=1}^{N_{mix}} \frac{1}{2N_{mix}} g_{ij}$, where g_{ij} is the sum of the two alleles, a_{1ij} and a_{2ij} , for the i th individual in the mixture ($i = 1, \dots, N_{mix}$). Hence,

$$\begin{aligned} \text{Var}(X_j) &= \text{Var}(q_j - p_j) = \sum_{i=1}^{N_{mix}} \text{Var}\left(\frac{1}{2N_{mix}} g_{ij}\right) \\ &= \left(\frac{1}{2N_{mix}}\right)^2 [\text{Var}(a_{11j}) + \text{Var}(a_{21j}) + \dots + \text{Var}(a_{1N_{mix}j}) + \text{Var}(a_{2N_{mix}j})] \\ &= \left(\frac{1}{2N_{mix}}\right)^2 \cdot 2N_{mix} \cdot [p_j(1 - p_j)] = \frac{p_j(1 - p_j)}{2N_{mix}} \end{aligned}$$

If individual i is *not* in the mixture, Y_{ij} and X_j are independent and, hence, $\text{Cov}(Y_{ij}, X_j|out) = 0$.

If individual i is in the mixture, Y_{ij} and X_j share the following elements:

$$\frac{1}{2}g_{ij} \text{ and } \frac{1}{2N_{mix}}g_{ij} \left[\text{or } \frac{a_{1ij}+a_{2ij}}{2} \text{ and } \frac{a_{1ij}+a_{2ij}}{2N_{mix}} \right].$$

Hence, the covariance between Y_{ij} and X_j , $\text{Cov}(Y_j, X_j|in)$ is:

$$\begin{aligned} \text{Cov}(Y_j, X_j|in) &= \text{Cov}\left(\frac{1}{2}g_{ij}, \frac{1}{2N_{mix}}g_{ij}\right) \\ &= \frac{1}{2} * \frac{1}{2N_{mix}} [Var(a_{1ij}) + Var(a_{2ij})] = \frac{p_j(1-p_j)}{2N_{mix}}. \end{aligned}$$

Assuming many SNPs are to be used in the test, the expectation of b_i can be defined by the expectation of the ratios:

$$E(b_i|in) = E\left[\frac{\text{Cov}(Y_{ij}, X_j|in)}{\text{Var}(X_j)}\right] = 1, \text{ and}$$

$E(b_i|out) = E\left[\frac{\text{Cov}(Y_{ij}, X_j|out)}{\text{Var}(X_j)}\right] = 0$ (i.e. by substituting in the corresponding elements outlined above).

Hence, if the individual of interest is in the mixture, the regression coefficient b_i has an expectation of 1, and if the individual of interest is not in the mixture, b_i has an expectation of 0.

The variance of b_i is defined as:

$$\text{Var}(b_i) = \sigma^2 [X'X]^{-1}, \text{ where } \sigma^2 \text{ is estimated by } \hat{\sigma}^2.$$

If individual i is not in the mixture, $\hat{\sigma}^2 = \text{Var}(Y_{ij}) = \frac{p_j(1-p_j)}{2}$ (derived above).

If individual i is in the mixture, $\hat{\sigma}^2 = \frac{(N_{mix}-1)}{N_{mix}} \cdot \frac{p_j(1-p_j)}{2}$.

Where $X'X = \sum_{j=1}^s (q_j - p_j)^2 = s * \text{Var}(X_j) = s * \frac{p_j(1-p_j)}{2N_{mix}}$,

the variance of b_i for an individual not in the mixture is

$$\text{Var}(b_i|\text{out}) = \hat{\sigma}^2|\text{out}. [X'X]^{-1} = \frac{\frac{p_j(1-p_j)}{2}}{s \cdot \frac{p_j(1-p_j)}{2N_{mix}}} = \frac{N_{mix}}{s},$$

and the variance of b_i for an individual in the mixture is

$$\text{Var}(b_i|\text{in}) = \hat{\sigma}^2|\text{in}. [X'X]^{-1} = \frac{\frac{(N_{mix}-1) p_j(1-p_j)}{2}}{s \cdot \frac{p_j(1-p_j)}{2N_{mix}}} = \frac{(N_{mix}-1)}{s}.$$

B.2.2 Population frequencies estimated:

In this scenario, the population frequencies, p_j , are not assumed to be known and, thus, they are estimated. These estimates of the population frequencies are denoted by \hat{p}_j , and are obtained from a weighted average of the corresponding allele frequency in the mixture and in the reference group:

$$\hat{p}_j = \frac{a_{1ij} + a_{2ij} + \dots + a_{1,N_{mix},j} + a_{2,N_{mix},j} + a_{1,(N_{mix}+1),j} + a_{2,(N_{mix}+1),j} + \dots + a_{1,(N_{mix}+N_{ref}),j} + a_{2,(N_{mix}+N_{ref}),j}}{2(N_{mix} + N_{ref})}$$

The regression here is again Y_{ij} on X_j , but Y_{ij} is now $(y_{ij} - \hat{p}_j)$ and X_j is now $(q_j - \hat{p}_j)$. As before, b_i is thus estimated by:

$$\hat{b}_i = \frac{\sum_{j=1}^s (y_{ij} - \hat{p}_j) (q_j - \hat{p}_j)}{\sum_{j=1}^s (q_j - \hat{p}_j)^2}.$$

In this scenario, $\text{Var}(y_{ij})$ and $\text{Var}(q_j)$ are derived as before, but here, \hat{p}_j is also a random variable, and has variance:

$$\begin{aligned}
 \text{Var}(\hat{p}_j) &= \sum_{i=1}^{(N_{\text{mix}}+N_{\text{ref}})} \text{Var} \left[\frac{1}{2(N_{\text{mix}}+N_{\text{ref}})} g_{ij} \right] \\
 &= \left[\frac{1}{2(N_{\text{mix}} + N_{\text{ref}})} \right]^2 \cdot [\text{Var}(a_{11j}) + \text{Var}(a_{21j}) + \dots + \text{Var}(a_{1, (N_{\text{mix}}+N_{\text{ref}}), j}) + \text{Var}(a_{2, (N_{\text{mix}}+N_{\text{ref}}), j})] \\
 &= \left[\frac{1}{2(N_{\text{mix}}+N_{\text{ref}})} \right]^2 \cdot 2(N_{\text{mix}} + N_{\text{ref}}) \cdot p_j(1 - p_j) = \frac{p_j(1-p_j)}{2(N_{\text{mix}}+N_{\text{ref}})}.
 \end{aligned}$$

Visscher *et al.* state that if the individual of interest is not in the mixture, $\text{Cov}(Y_{ij}, X_{j|\text{out}})$ is again 0. However, this ignores the possibility that an individual could be in the reference group; in this situation, the covariance, $\text{Cov}(Y_{ij}, X_{j|\text{in ref}})$ will be non-zero. The stated covariance, $\text{Cov}(Y_{ij}, X_{j|\text{out}}) = 0$ thus applies only to individuals who are in neither of the two test groups. In order to derive $\text{Cov}(Y_j, X_{j|\text{in mix}})$ and $\text{Cov}(Y_j, X_{j|\text{in ref}})$ it is first necessary to derive the variance components between y_{ij} and q_j ; y_{ij} and \hat{p}_j ; and q_j and \hat{p}_j :

$$\begin{aligned}
 \text{Cov}(y_{ij} - \hat{p}_j, q_j - \hat{p}_j) &= \text{Cov}(aX + bY, cW + dV) \\
 &= ac \text{Cov}(X,W) + ad \text{Cov}(X,V) + bc \text{Cov}(Y,W) + bd \text{Cov}(Y,V) \\
 &= \text{Cov}(y_{ij}, q_j) - \text{Cov}(y_{ij}, \hat{p}_j) - \text{Cov}(q_j, \hat{p}_j) + \text{Var}(\hat{p}_j) \tag{1}
 \end{aligned}$$

For an individual in the reference group, the covariance $\text{Cov}(y_{ij}, q_j | \text{in ref}) = 0$.

(2a)

For an individual in the mixture, y_{ij} and q_j have the following shared elements:

$$\frac{a_{1ij}+a_{2ij}}{2} \text{ and } \frac{a_{1ij}+a_{2ij}}{2N_{mix}},$$

$$\text{and, hence, } \text{Cov}(y_{ij}, q_j | \text{in mix}) = \frac{1}{2} \cdot \frac{1}{2N_{mix}} \cdot [\text{Var}(a_{1ij}) + \text{Var}(a_{2ij})] = \frac{p_j(1-p_j)}{2N_{mix}}. \quad (2b)$$

Regardless of whether the individual is in the mixture or the reference group, y_{ij}

and \hat{p}_j have the following shared elements: $\frac{a_{1ij}+a_{2ij}}{2}$ and $\frac{a_{1ij}+a_{2ij}}{2(N_{mix}+N_{ref})}$.

$$\text{Hence, } \text{Cov}(y_{ij}, \hat{p}_j) = \frac{1}{2} \cdot \frac{1}{2(N_{mix}+N_{ref})} \cdot [\text{Var}(a_{1ij}) + \text{Var}(a_{2ij})] = \frac{p_j(1-p_j)}{2(N_{mix}+N_{ref})}. \quad (3)$$

Similarly, q_j and \hat{p}_j have the following shared elements:

$$\frac{a_{11j}+a_{21j}+\dots+a_{1,N_{mix},j}+a_{2,N_{mix},j}}{2N_{mix}} \text{ and } \frac{a_{11j}+a_{21j}+\dots+a_{1,N_{mix},j}+a_{2,N_{mix},j}}{2(N_{mix}+N_{ref})}.$$

Thus, $\text{Cov}(q_j, \hat{p}_j) =$

$$\begin{aligned} & \frac{1}{2N_{mix}} \cdot \frac{1}{2(N_{mix} + N_{ref})} \cdot [\text{Var}(a_{11j}) + \text{Var}(a_{21j}) + \dots + \text{Var}(a_{1,N_{mix},j}) \\ & \quad + \text{Var}(a_{2,N_{mix},j})] \\ & = \frac{p_j(1-p_j)}{2(N_{mix}+N_{ref})} \end{aligned} \quad (4)$$

$$\text{From the previous page, } \text{Var}(\hat{p}_j) = \frac{p_j(1-p_j)}{2(N_{mix}+N_{ref})}. \quad (5)$$

Hence, using the formula given in (1), and using elements (2) – (5),

$$\text{Cov}(Y_j, X_j | \text{in mix}) = \left[\frac{1}{2N_{mix}} - \frac{1}{2(N_{mix}+N_{ref})} - \frac{1}{2(N_{mix}+N_{ref})} + \frac{1}{2(N_{mix}+N_{ref})} \right] \cdot [p_j(1-p_j)]$$

$$= \frac{1}{2} \left[\frac{1}{N_{mix}} - \frac{1}{(N_{mix} + N_{ref})} \right] [p_j(1 - p_j)], \text{ and}$$

$$\begin{aligned} \text{Cov}(Y_j, X_j | \text{in ref}) &= \left[\frac{1}{2(N_{mix} + N_{ref})} - \frac{1}{2(N_{mix} + N_{ref})} + \frac{1}{2(N_{mix} + N_{ref})} \right] \cdot [p_j(1 - p_j)] \\ &= -\frac{1}{2} \left[\frac{1}{(N_{mix} + N_{ref})} \right] [p_j(1 - p_j)]. \end{aligned}$$

The variance of X_j can now also be derived using the above variance components:

$$\text{Var}(X_j) = \text{Var}(q_j) + \text{Var}(-\hat{p}_j) - 2 \cdot \text{Cov}(q_j, -\hat{p}_j).$$

As above, $\text{Var}(q_j) = \frac{p_j(1-p_j)}{2N_{mix}}$; $\text{Var}(-\hat{p}_j)$ is presented in equation (5); and $\text{Cov}(q_j, -\hat{p}_j)$ is presented in equation (4).

$$\begin{aligned} \text{Hence, } \text{Var}(X_j) &= \left[\frac{1}{N_{mix}} + \frac{1}{(N_{mix} + N_{ref})} - \frac{2}{(N_{mix} + N_{ref})} \right] [p_j(1 - p_j)] \\ &= \frac{1}{2} \left[\frac{1}{N_{mix}} - \frac{1}{(N_{mix} + N_{ref})} \right] [p_j(1 - p_j)]. \quad (6) \end{aligned}$$

The expectation of b_j can now be derived for individuals in the mixture (*in mix.*), in the reference group (*in ref.*), and for individuals in neither group (*out*):

$$E(b_j | \text{in mix.}) = E \left[\frac{\text{Cov}(Y_j, X_j | \text{in})}{\text{Var}(X_j)} \right] = 1;$$

$$E(b_j | \text{in ref.}) = \frac{-\frac{1}{2} \left[\frac{1}{(N_{mix} + N_{ref})} \right] [p_j(1 - p_j)]}{\frac{1}{2} \left[\frac{1}{N_{mix}} - \frac{1}{(N_{mix} + N_{ref})} \right] [p_j(1 - p_j)]} = -\frac{N_{mix}}{N_{ref}}; \text{ and}$$

$$E(b_j | \text{out}) = 0.$$

In order to derive the variance of b_i , the general formula $\text{Var}(\hat{b}) = \sigma^2[X'X]^{-1}$ is again used.

As before, σ^2 is estimated by $\hat{\sigma}^2$, where

$$\hat{\sigma}^2|_{\text{out}} = \text{Var}(y_{ij}) = \frac{p_j(1-p_j)}{2},$$

$$\hat{\sigma}^2|_{\text{in mix.}} = \frac{(N_{\text{mix}}-1) p_j(1-p_j)}{N_{\text{mix}} 2}, \text{ and}$$

$$\hat{\sigma}^2|_{\text{in ref.}} = \frac{(N_{\text{ref}}-1) p_j(1-p_j)}{N_{\text{ref}} 2}$$

As before, $X'X = \sum_{j=1}^s (q_j - p_j)^2 = s * \text{Var}(Q_j)$.

$$\text{From (6), } \text{Var}(Q_j) = \frac{1}{2} \left[\frac{1}{N_{\text{mix}}} - \frac{1}{(N_{\text{mix}}+N_{\text{ref}})} \right] [p_j(1-p_j)],$$

$$\text{Hence, } X'X = \frac{s}{2} \left[\frac{1}{N_{\text{mix}}} - \frac{1}{(N_{\text{mix}}+N_{\text{ref}})} \right] [p_j(1-p_j)].$$

$$\text{Thus, } \text{Var}(\hat{b}|_{\text{out}}) = \frac{\frac{p_j(1-p_j)}{2}}{\frac{s}{2} \left[\frac{1}{N_{\text{mix}}} - \frac{1}{(N_{\text{mix}}+N_{\text{ref}})} \right] [p_j(1-p_j)]},$$

which simplifies to

$$\text{Var}(\hat{b}|_{\text{out}}) = \frac{N_{\text{mix}}}{s} \cdot \frac{(N_{\text{mix}}+N_{\text{ref}})}{N_{\text{ref}}}, \text{ and}$$

$$\text{Var}(\hat{b}|_{\text{in mix}}) = \frac{\frac{(N_{\text{mix}}-1) p_j(1-p_j)}{N_{\text{mix}} 2}}{\frac{s}{2} \left[\frac{1}{N_{\text{mix}}} - \frac{1}{(N_{\text{mix}}+N_{\text{ref}})} \right] [p_j(1-p_j)]} = \frac{(N_{\text{mix}}-1)}{s} \cdot \frac{(N_{\text{mix}}+N_{\text{ref}})}{N_{\text{ref}}}.$$

B.2.3 Derivation of the Test Statistic:

Visscher & Hill propose a statistic, T , to test the null hypothesis that the individual of interest is in the mixture [$E(b_i) = 1$] against an alternative hypothesis that the individual is not in the mixture ($b < 1$):

$$T = \frac{(X - \mu_0)^2}{\sigma^2} = \frac{(\hat{b} - 1)^2}{\text{var}(b|\text{out})} \sim \chi^2_{1 \text{ d.f.}}$$

Alternatively, to test the null hypothesis that the individual of interest is not in the mixture [$E(b=0)$] against the alternative hypothesis that he/she is in the mixture ($b > 0$), Visscher & Hill propose the statistic:

$$T = \frac{(\hat{b} - 1)^2}{\text{var}(b|\text{out})} \sim \text{non-central } \chi^2_{1 \text{ d.f.}, \lambda}$$

$$\text{where } \lambda = \left(\frac{\mu}{\sigma}\right)^2 = \left[\frac{1}{\text{var}(\hat{b}|\text{out})}\right]^2 = (s/N_{\text{mix}})[N_{\text{ref}}/(N_{\text{mix}} + N_{\text{ref}})].$$

As Section 2.2.3 discusses, however, the situation where two groups are compared (such as in a case-control study) has three possible outcomes (e.g. case, control or neither group), and a two-tailed hypothesis test therefore seems more appropriate than the one-tailed tests outlined above. Noting that the null hypothesis that an individual is not in a study will usually be more useful than its reverse (i.e. assuming that the individual *is* in the study under the null), a two-tailed hypothesis test can be expressed as:

$$Z = \frac{X - \mu_0}{\sigma} = \frac{\hat{b}}{\sqrt{\text{var}(b|\text{out})}} \sim N(0, 1^2).$$

This Z test is used throughout in preference to the test statistics proposed by Visscher & Hill.

Appendix C.

This section contains R code for simulating the datasets described in sections 3.3.1.1 and 3.3.2.1, and for performing the analyses described in sections 3.3.1.2 and 3.3.2.2. All R code may be cut and pasted directly into R to replicate any of the procedures. Model output is also provided to show both interim and overall results from each analysis.

C.1. Scenario 1

C.1.1 R code for simulating the data

#set up data structure

```
set.seed(18984)
numsubs.study<-c(1000,2000,3000,4000,2500,2500)
numsubs<-sum(numsubs.study)
numstudies<-length(numsubs.study)
study.id<-rep(1:numstudies,numsubs.study)
```

#set up model structure and parameters

```
numpara<-3
beta0<-125
betaAGE<-0.25
betaSNP<-0.5
MAF<-0.2
```

#simulate data

```
AGE<-runif(numsubs,50,70) - 60
SNP.1<-rbinom(numsubs,1,MAF)
SNP.2<-rbinom(numsubs,1,MAF)
SNP<-SNP.1+SNP.2
lp<-beta0+betaAGE*AGE+betaSNP*SNP
SBP<-rnorm(numsubs,lp,11)
```

C.1.2 R Code & Output for Analysis 1 (ILMA):

#Analyse all studies together

```
model.overall<-lm(SBP~AGE+SNP)
summary(model.overall)
```

Abbreviated output from overall regression analysis:

OVERALL ANALYSIS

Call:

```
lm(formula = SBP ~ AGE + SNP)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.15770	0.10943	1143.724	<2e-16
AGE	0.25937	0.01549	16.745	<2e-16
SNP	0.44796	0.15806	2.834	0.0046

C.1.3 R code & Output for Analysis 2 (DataSHIELD analysis):

```
#####
```

```
# STUDY SPECIFIC ANALYSES #
```

```
#####
```

#create empty results matrices

```
beta.s<-matrix(NA,nrow=numpara,ncol=numstudies)
se.s<-matrix(NA,nrow=numpara,ncol=numstudies)
```

#work with each study one at a time

```
for(k in 1:numstudies)
{
  SBP.s<-SBP[study.id==k]
  AGE.s<-AGE[study.id==k]
  SNP.s<-SNP[study.id==k]
```

```

model.study.specific<-lm(SBP.s~AGE.s+SNP.s)
print(summary(model.study.specific))
beta.s[,k]<-summary(model.study.specific)$coefficients[,1]
se.s[,k]<-summary(model.study.specific)$coefficients[,2]
}

```

Abbreviated output from study-specific analyses:

STUDY 1

Call:

```
lm(formula = SBP.s ~ AGE.s + SNP.s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	124.95070	0.43910	284.560	< 2e-16
AGE.s	0.31155	0.06315	4.933	9.47e-07
SNP.s	1.66853	0.67835	2.460	0.0141

STUDY 2

Call:

```
lm(formula = SBP.s ~ AGE.s + SNP.s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.29081	0.29894	419.119	< 2e-16
AGE.s	0.22046	0.04265	5.169	2.59e-07
SNP.s	-0.28092	0.42458	-0.662	0.508

STUDY 3

Call:

```
lm(formula = SBP.s ~ AGE.s + SNP.s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.1460	0.2467	507.308	<2e-16
AGE.s	0.2990	0.0349	8.566	<2e-16
SNP.s	0.8101	0.3483	2.326	0.0201

STUDY 4

Call:

```
lm(formula = SBP.s ~ AGE.s + SNP.s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.10805	0.21473	582.633	<2e-16
AGE.s	0.27995	0.03003	9.321	<2e-16
SNP.s	0.44297	0.30450	1.455	0.146

STUDY 5

Call:

```
lm(formula = SBP.s ~ AGE.s + SNP.s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.23344	0.26195	478.086	< 2e-16
AGE.s	0.24625	0.03726	6.609	4.7e-11
SNP.s	0.37170	0.38534	0.965	0.335

STUDY 6

Call:

```
lm(formula = SBP.s ~ AGE.s + SNP.s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.12993	0.26289	475.974	< 2e-16
AGE.s	0.20305	0.03727	5.448	5.59e-08
SNP.s	0.25418	0.39068	0.651	0.515

#####

META-ANALYSIS

#####

#set up analysis weights

analysis.wt<-numsubs.study/numsubs

#set up a vector of 1s to use in summing precisions

simple.sum<-rep(1,numstudies)

#calculate mean of regression coefficients weighted for study sample sizes

“%*%” denotes vector multiplication

beta.overall<-beta.s%*%analysis.wt

#convert standard errors into precisions

precision.s <-1/(se.s)^2

#sum precisions across studies

precision.overall<-precision.s%*%simple.sum

#convert precisions back to standard errors

```
se.overall <- 1/(precision.overall)^0.5
```

#round outputs

```
beta.overall<-round(beta.overall,digits=4)
```

```
se.overall<-round(se.overall,digits=4)
```

#create output results matrix

```
meta.analysis.results<-cbind(beta.overall,se.overall)
```

```
dimnames(meta.analysis.results)<-
```

```
list(c("Intercept","AGE","SNP"),c("Coefficients","SE"))
```

#print output

```
print(analysis.wt)
```

```
print(meta.analysis.results)
```

Output from meta-analysis:

```
print (analysis.wt)
```

```
[1] 0.06667 0.13333 0.20000 0.26667 0.16667 0.16667
```

```
print (meta.analysis.results)
```

```
> print(meta.analysis.results)
```

	Coefficients	SE
Intercept	125.1541	0.1094
AGE	0.2595	0.0155
SNP	0.4582	0.1580

C.2. Scenario 2

C.2.1 R code for simulating the data

```
#####
# SIMULATION #
#####

#Start R code preparation
#First maximise memory allocation
memory.limit(4095)

#For convenience, start by setting up file names ahead of time
DC1.data.file<-"C:/DataSHIELD.Example/DC1/Study.1.csv"
DC2.data.file<-"C:/DataSHIELD.Example/DC2/Study.2.csv"
DC3.data.file<-"C:/DataSHIELD.Example/DC3/Study.3.csv"
DC4.data.file<-"C:/DataSHIELD.Example/DC4/Study.4.csv"
DC5.data.file<-"C:/DataSHIELD.Example/DC5/Study.5.csv"
DC6.data.file<-"C:/DataSHIELD.Example/DC6/Study.6.csv"
AC.beta.vector<-"C:/DataSHIELD.Example/AC/beta.vector.csv"
ALL.data.file<-"C:/DataSHIELD.Example/Study.ALL.csv"

#SET UP DATA STRUCTURE

#Random number seed so results can be replicated
```

```
set.seed(1028)
```

#Specify study sizes and generate IDs for studies and individuals

```
numsubs.study<-c(2000,3000,1500,300,2000,700)
```

```
numsubs<-sum(numsubs.study)
```

```
numstudies<-length(numsubs.study)
```

```
study.id<-rep(1:numstudies,numsubs.study)
```

```
id<-c(1:numsubs.study[1], 1:numsubs.study[2], 1:numsubs.study[3],
```

```
1:numsubs.study[4], 1:numsubs.study[5], 1:numsubs.study[6])
```

#SET UP MODEL STRUCTURE AND PARAMETERS

#Number of and values of regression coefficients

```
numpara<-4
```

```
beta0<--0.3
```

```
beta.bmi<-0.02
```

```
beta.bmi456<-0.04
```

```
beta.snp<-0.5
```

#Minor allele frequency

```
MAF<-0.3
```

#SIMULATE DATA

#Generate covariates

```
bmi<- rnorm(numsubs,mean=23,sd=4)-23
```

```
bmi456<-c(rep(0,6500),bmi[6501:9500])
```

```
snp<-rbinom(numsubs,2,MAF)
```

#Generate linear predictor and equivalent probabilities of response

```
lp<-beta0 + beta.bmi*bmi +beta.bmi456*bmi456 + beta.snp*snp
```

```
probresp<-exp(lp)/(1+exp(lp))
```

#Randomly sample case control status

```
CC<-rbinom(numsubs,1,probresp)
```

#ASSEMBLE AND WRITE OUT COMPLETE DATA SET

```
all.data<-data.frame(study.id,id,CC,bmi,snp,bmi456)
```

```
write.csv(all.data,file=ALL.data.file,row.names=FALSE)
```

#PREPARE AND WRITE OUT DATA FILES FOR EACH STUDY INDIVIDUALLY

```
Study<-list()
```

```
Study[[1]]<-all.data[study.id==1,]
```

```
write.csv(Study[[1]],file=DC1.data.file,row.names=FALSE)
```

```
Study[[2]]<-all.data[study.id==2,]
```

```
write.csv(Study[[2]],file=DC2.data.file,row.names=FALSE)
```

```
Study[[3]]<-all.data[study.id==3,]
```

```
write.csv(Study[[3]],file=DC3.data.file,row.names=FALSE)
```

```
Study[[4]]<-all.data[study.id==4,]

write.csv(Study[[4]],file=DC4.data.file,row.names=FALSE)

Study[[5]]<-all.data[study.id==5,]

write.csv(Study[[5]],file=DC5.data.file,row.names=FALSE)

Study[[6]]<-all.data[study.id==6,]

write.csv(Study[[6]],file=DC6.data.file,row.names=FALSE)
```

C.2.2 R Code & Output for Analysis 1 (ILMA):

#Fit model on all data sets combined

```
ALL.data.file<-"C:/DataSHIELD.Example/Study.ALL.csv"
ALL.data<-read.table(file=ALL.data.file, sep="," ,header=T)

summary(glm(CC~bmi+ bmi456 + snp,family=binomial(logit),data=ALL.data))
```

Output from the overall analysis:

Coefficients:

	Estimate	Std Error	Z value	Pr(> z)
(Intercept)	-0.32956	0.02838	-11.612	<2e-16 ***
BMI	0.02300	0.00621	3.703	0.00021 *** 3
BMI.456	0.04126	0.01140	3.620	0.00029 *** 5
SNP	0.55173	0.03295	16.746	< 2e-16 ***

Residual deviance: 12825 on 9496 degrees of freedom

C.2.3 R Code to perform Analysis 2 (DataSHIELD analysis):

#####

R CODE TO SET UP ANALYSIS

#####

#Specify folder for storing objects on the AC

AC.Directory<-"C:/DataSHIELD.Example/AC/"

#Create initial vector of regression coefficients for first iteration

beta.vect.next<-c(0,0,0,0)

#Save to the folder where data computers can find it

save(beta.vect.next,file=paste(AC.Directory,"beta.vect.next.RData",sep=""))

#Iterations need to be counted. Start off with the count at 0

#and increment by 1 at each new iteration

```
iteration.count<-0
```

#Provide arbitrary starting value for deviance to enable subsequent calculation of the

#change in deviance between iterations

```
dev.old<-9.99e+99
```

#Convergence state needs to be monitored. Start by allocating

#a "convergence not met" status

```
converge.state<-"NOT MET"
```

#Define a convergence criterion. This value of epsilon corresponds to that used

#by default for GLMs in R (see section S3 for details)

```
epsilon<-1.0e-08
```

```
#####
```

```
# R CODE TO CARRY OUT A PARTITIONED IRLS FIT ONE ITERATION AT A TIME #
```

```
# RUN THIS WHOLE BLOCK OF CODE ONE ITERATION AT A TIME UNTIL THE #
```

```
# MODEL OUTPUT INDICATES THAT CONVERGENCE HAS BEEN ACHIEVED
```

```
#####
```

#Increment count of iterations

```
iteration.count<-iteration.count+1
```

```
#R CODE THAT WOULD RUN LOCALLY ON EACH OF THE REMOTE DATA COMPUTERS
```

```
#####
```

```
#START STUDY 1
```

#Read in full data

```
data.DC<-read.table(file=DC1.data.file, sep="," ,header=T)
```

```

#Strip out the first column (which is a duplicate index column)
data.DC<-data.DC[,-1]

#Calculate number of subjects available in the current study
#(by enumerating length of ID column)
nsubs<-length(data.DC$id)

#Define the design matrix (matrix of covariates) to contain BMI, SNP and the
#interaction covariate and add a column of 1s at the start for the regression
constant
X.mat<-cbind(rep(1,nsubs),data.DC$bmi,data.DC$bmi456,data.DC$snp)

#Load the current value of the beta vector (vector of regression coefficients) from
its
#location on the AC
load(file=paste(AC.Directory,"beta.vect.next.RData",sep=""))

# Use this current value of the beta vector to calculate elements from the current
study
beta.vect<-beta.vect.next

# Calculate linear predictors from observed covariate values and elements of the
# current beta vector
lp.current<-beta.vect[1]+beta.vect[2]*X.mat[,2]+beta.vect[3]*X.mat[,3]+
beta.vect[4]*X.mat[,4]

# Apply inverse logistic transformation
mu.current<-exp(lp.current)/(1+exp(lp.current))

# Derive variance function and diagonal elements for weight matrix (using
squared
# first differential of link function)
var.i<-(mu.current*(1-mu.current))
g2.i<-(1/(mu.current*(1-mu.current)))^2
W.mat<-diag(1/(var.i*g2.i))

```

#Calculate information matrix

```
info.matrix<-t(X.mat)%*%W.mat%*%X.mat
```

#Derive u terms for score vector

```
u.i<- (data.DC$CC-mu.current)* (1/(mu.current*(1-mu.current)))
```

#Calculate score vector

```
score.vect<-t(X.mat)%*%W.mat%*%u.i
```

#Calculate log likelihood and deviance contribution for current study

#For convenience, ignore the element of deviance that relates to the full saturated # model, because that will cancel out in calculating the change in deviance from one

iteration to the next (Dev.total - Dev.old [see below]) because the element relating

to the saturated model will be the same at every iteration).

```
log.L<-sum(data.DC$CC*log(mu.current) + (1-data.DC$CC)*log(1-mu.current))
```

```
dev<- -2*log.L
```

#Create study specific versions of all key model components

```
info.matrix.1<-info.matrix
```

```
score.vect.1<-score.vect
```

```
dev.1<-dev
```

```
nsubs.1<-nsubs
```

#Send all the key model components from the current study to the AC

```
save(info.matrix.1,file=paste(AC.Directory,"info.matrix.1.RData",sep=""))
```

```
save(score.vect.1,file=paste(AC.Directory,"score.vect.1.RData",sep=""))
```

```
save(dev.1,file=paste(AC.Directory,"dev.1.RData",sep=""))
```

```
save(nsubs.1,file=paste(AC.Directory,"nsubs.1.RData",sep=""))
```

#END STUDY 1

```
#####
```

```
#####  
#START STUDY 2  
  
#Read in full data  
data.DC<-read.table(file=DC2.data.file, sep=",",header=T)  
  
#Strip out first column  
data.DC<-data.DC[,-1]  
  
#Calculate number of subjects available in the current study  
 #(by enumerating length of ID column)  
nsubs<-length(data.DC$id)  
  
#Define design matrix (matrix of covariates) to contain BMI, SNP and the  
#interaction covariate and add a column of 1s at the start for the regression  
constant  
X.mat<-cbind(rep(1,nsubs),data.DC$bmi,data.DC$bmi456,data.DC$snp)  
  
#Load the current value of the beta vector (vector of regression coefficients) from  
its  
#location on the AC computer  
load(file=paste(AC.Directory,"beta.vect.next.RData",sep=""))  
  
# Use this current value of the beta vector to calculate elements from the current  
study  
beta.vect<-beta.vect.next  
  
# Calculate linear predictors from observed covariate values and elements of  
# current beta vector  
lp.current<-beta.vect[1]+beta.vect[2]*X.mat[,2]+beta.vect[3]*X.mat[,3]+  
beta.vect[4]*X.mat[,4]  
  
# Apply inverse logistic transformation  
mu.current<-exp(lp.current)/(1+exp(lp.current))
```

```
# Derive variance function and diagonal elements for weight matrix (using squared
```

```
# first differential of link function)
```

```
var.i<-(mu.current*(1-mu.current))
```

```
g2.i<-(1/(mu.current*(1-mu.current)))^2
```

```
W.mat<-diag(1/(var.i*g2.i))
```

```
#Calculate information matrix
```

```
info.matrix<-t(X.mat)%*%W.mat%*%X.mat
```

```
#Derive u terms for score vector
```

```
u.i<- (data.DC$CC-mu.current)* (1/(mu.current*(1-mu.current)))
```

```
#Calculate score vector
```

```
score.vect<-t(X.mat)%*%W.mat%*%u.i
```

```
#Calculate log likelihood and deviance contribution for current study
```

```
#For convenience, ignore the element of deviance that relates to the full saturated  
# model, because that will cancel out in calculating the change in deviance from  
one
```

```
# iteration to the next (Dev.total - Dev.old [see below]) because the element  
relating
```

```
# to the saturated model will be the same at every iteration).
```

```
log.L<-sum(data.DC$CC*log(mu.current) + (1-data.DC$CC)*log(1-mu.current))
```

```
dev<- -2*log.L
```

```
#Create study specific versions of all key model components
```

```
info.matrix.2<-info.matrix
```

```
score.vect.2<-score.vect
```

```
dev.2<-dev
```

```
nsubs.2<-nsubs
```

```
#Send all of the key model components from the current study to the AC
```

```
save(info.matrix.2,file=paste(AC.Directory,"info.matrix.2.RData",sep=""))
```

```
save(score.vect.2,file=paste(AC.Directory,"score.vect.2.RData",sep=""))
```

```
save(dev.2,file=paste(AC.Directory,"dev.2.RData",sep=""))
save(nsubs.2,file=paste(AC.Directory,"nsubs.2.RData",sep=""))
```

```
#END STUDY 2
```

```
#####
```

```
#####
```

```
#START STUDY 3
```

```
#Read in full data
```

```
data.DC<-read.table(file=DC3.data.file, sep=",",header=T)
```

```
#Strip out first column
```

```
data.DC<-data.DC[,-1]
```

```
#Calculate number of subjects available in the current study
```

```
 #(by enumerating length of ID column)
```

```
nsubs<-length(data.DC$id)
```

```
#Define design matrix (matrix of covariates) to contain BMI, SNP and the
```

```
#interaction covariate and add a column of 1s at the start for the regression  
constant
```

```
X.mat<-cbind(rep(1,nsubs),data.DC$bmi,data.DC$bmi456,data.DC$snp)
```

```
#Load the current value of the beta vector (vector of regression coefficients) from  
its
```

```
#location on the AC computer (stored during activation of block 2 of R code)
```

```
load(file=paste(AC.Directory,"beta.vect.next.RData",sep=""))
```

```
# Use this current value of the beta vector to calculate elements from the current  
study
```

```
beta.vect<-beta.vect.next
```

```
# Calculate linear predictors from observed covariate values and elements of
```

```
# current beta vector
```

```
lp.current<-beta.vect[1]+beta.vect[2]*X.mat[,2]+beta.vect[3]*X.mat[,3]+
beta.vect[4]*X.mat[,4]
```

Apply inverse logistic transformation

```
mu.current<-exp(lp.current)/(1+exp(lp.current))
```

Derive variance function and diagonal elements for weight matrix (using squared

first differential of link function)

```
var.i<-(mu.current*(1-mu.current))
```

```
g2.i<-(1/(mu.current*(1-mu.current)))^2
```

```
W.mat<-diag(1/(var.i*g2.i))
```

#Calculate information matrix

```
info.matrix<-t(X.mat)%*%W.mat%*%X.mat
```

#Derive u terms for score vector

```
u.i<- (data.DC$CC-mu.current)* (1/(mu.current*(1-mu.current)))
```

#Calculate score vector

```
score.vect<-t(X.mat)%*%W.mat%*%u.i
```

#Calculate log likelihood and deviance contribution for current study

**#For convenience, ignore the element of deviance that relates to the full saturated
model, because that will cancel out in calculating the change in deviance from
one**

**# iteration to the next (Dev.total – Dev.old [see below]) because the element
relating**

to the saturated model will be the same at every iteration).

```
log.L<-sum(data.DC$CC*log(mu.current) + (1-data.DC$CC)*log(1-mu.current))
```

```
dev<- -2*log.L
```

#Create study specific versions of all key model components

```
info.matrix.3<-info.matrix
```

```
score.vect.3<-score.vect
dev.3<-dev
nsubs.3<-nsubs
```

#Send all of the key model components from the current study to the AC

```
save(info.matrix.3,file=paste(AC.Directory,"info.matrix.3.RData",sep=""))
save(score.vect.3,file=paste(AC.Directory,"score.vect.3.RData",sep=""))
save(dev.3,file=paste(AC.Directory,"dev.3.RData",sep=""))
save(nsubs.3,file=paste(AC.Directory,"nsubs.3.RData",sep=""))
```

#END STUDY 3

```
#####
```

```
#####
```

#START STUDY 4

#Read in full data

```
data.DC<-read.table(file=DC4.data.file, sep=",",header=T)
```

#Strip out first column

```
data.DC<-data.DC[,-1]
```

#Calculate number of subjects available in the current study

#(by enumerating length of ID column)

```
nsubs<-length(data.DC$id)
```

#Define design matrix (matrix of covariates) to contain BMI, SNP and the

#interaction covariate and add a column of 1s at the start for the regression constant

```
X.mat<-cbind(rep(1,nsubs),data.DC$bmi,data.DC$bmi456,data.DC$snp)
```

#Load the current value of the beta vector (vector of regression coefficients) from its

#location on the AC computer

```
load(file=paste(AC.Directory,"beta.vect.next.RData",sep=""))
```

Use this current value of the beta vector to calculate elements from the current study

```
beta.vect<-beta.vect.next
```

Calculate linear predictors from observed covariate values and elements of current beta vector

```
lp.current<-beta.vect[1]+beta.vect[2]*X.mat[,2]+beta.vect[3]*X.mat[,3]+
beta.vect[4]*X.mat[,4]
```

Apply inverse logistic transformation

```
mu.current<-exp(lp.current)/(1+exp(lp.current))
```

Derive variance function and diagonal elements for weight matrix (using squared

first differential of link function)

```
var.i<-(mu.current*(1-mu.current))
g2.i<-(1/(mu.current*(1-mu.current)))^2
W.mat<-diag(1/(var.i*g2.i))
```

#Calculate information matrix

```
info.matrix<-t(X.mat)%*%W.mat%*%X.mat
```

#Derive u terms for score vector

```
u.i<- (data.DC$CC-mu.current)* (1/(mu.current*(1-mu.current)))
```

#Calculate score vector

```
score.vect<-t(X.mat)%*%W.mat%*%u.i
```

#Calculate log likelihood and deviance contribution for current study

#For convenience, ignore the element of deviance that relates to the full saturated model, because that will cancel out in calculating the change in deviance from one

iteration to the next (Dev.total - Dev.old [see below]) because the element relating

to the saturated model will be the same at every iteration).

```
log.L<-sum(data.DC$CC*log(mu.current) + (1-data.DC$CC)*log(1-mu.current))
dev<- -2*log.L
```

#Create study specific versions of all key model components

```
info.matrix.4<-info.matrix
score.vect.4<-score.vect
dev.4<-dev
nsubs.4<-nsubs
```

#Send all of the key model components from the current study to the AC

```
save(info.matrix.4,file=paste(AC.Directory,"info.matrix.4.RData",sep=""))
save(score.vect.4,file=paste(AC.Directory,"score.vect.4.RData",sep=""))
save(dev.4,file=paste(AC.Directory,"dev.4.RData",sep=""))
save(nsubs.4,file=paste(AC.Directory,"nsubs.4.RData",sep=""))
```

#END STUDY 4

```
#####
```

```
#####
```

#START STUDY 5

#Read in full data

```
data.DC<-read.table(file=DC5.data.file, sep=",",header=T)
```

#Strip out first column

```
data.DC<-data.DC[,-1]
```

#Calculate number of subjects available in the current study

#(by enumerating length of ID column)

```
nsubs<-length(data.DC$id)
```

#Define design matrix (matrix of covariates) to contain BMI, SNP and the

#interaction covariate and add a column of 1s at the start for the regression constant

```
X.mat<-cbind(rep(1,nsubs),data.DC$bmi,data.DC$bmi456,data.DC$snp)
```

#Load the current value of the beta vector (vector of regression coefficients) from its

#location on the AC computer

```
load(file=paste(AC.Directory,"beta.vect.next.RData",sep=""))
```

Use this current value of the beta vector to calculate elements from the current study

```
beta.vect<-beta.vect.next
```

Calculate linear predictors from observed covariate values and elements of current beta vector

```
lp.current<-beta.vect[1]+beta.vect[2]*X.mat[2]+beta.vect[3]*X.mat[3]+
beta.vect[4]*X.mat[4]
```

Apply inverse logistic transformation

```
mu.current<-exp(lp.current)/(1+exp(lp.current))
```

Derive variance function and diagonal elements for weight matrix (using squared

first differential of link function)

```
var.i<-(mu.current*(1-mu.current))
```

```
g2.i<-(1/(mu.current*(1-mu.current)))^2
```

```
W.mat<-diag(1/(var.i*g2.i))
```

#Calculate information matrix

```
info.matrix<-t(X.mat)%*%W.mat%*%X.mat
```

#Derive u terms for score vector

```
u.i<- (data.DC$CC-mu.current)* (1/(mu.current*(1-mu.current)))
```

#Calculate score vector

```
score.vect<-t(X.mat)%*%W.mat%*%u.i
```

#Calculate log likelihood and deviance contribution for current study

#For convenience, ignore the element of deviance that relates to the full saturated

model, because that will cancel out in calculating the change in deviance from one

iteration to the next (Dev.total - Dev.old [see below]) because the element relating

to the saturated model will be the same at every iteration).

```
log.L<-sum(data.DC$CC*log(mu.current) + (1-data.DC$CC)*log(1-mu.current))
```

```
dev<- -2*log.L
```

#Create study specific versions of all key model components

```
info.matrix.5<-info.matrix
```

```
score.vect.5<-score.vect
```

```
dev.5<-dev
```

```
nsubs.5<-nsubs
```

#Send all of the key model components from the current study to the AC

```
save(info.matrix.5,file=paste(AC.Directory,"info.matrix.5.RData",sep=""))
```

```
save(score.vect.5,file=paste(AC.Directory,"score.vect.5.RData",sep=""))
```

```
save(dev.5,file=paste(AC.Directory,"dev.5.RData",sep=""))
```

```
save(nsubs.5,file=paste(AC.Directory,"nsubs.5.RData",sep=""))
```

#END STUDY 5

#####

#####

#START STUDY 6

#Read in full data

```
data.DC<-read.table(file=DC6.data.file, sep=",",header=T)
```

#Strip out first column

```
data.DC<-data.DC[,-1]
```

#Calculate number of subjects available in the current study

#(by enumerating length of ID column)

```

nsubs<-length(data.DC$id)

#Define design matrix (matrix of covariates) to contain BMI, SNP and the
#interaction covariate and add a column of 1s at the start for the regression
constant
X.mat<-cbind(rep(1,nsubs),data.DC$bmi,data.DC$bmi456,data.DC$snp)

#Load the current value of the beta vector (vector of regression coefficients) from
its
#location on the AC computer (stored during activation of block 2 of R code)
load(file=paste(AC.Directory,"beta.vect.next.RData",sep=""))

# Use this current value of the beta vector to calculate elements from the current
study
beta.vect<-beta.vect.next

# Calculate linear predictors from observed covariate values and elements of
# current beta vector
lp.current<-beta.vect[1]+beta.vect[2]*X.mat[,2]+beta.vect[3]*X.mat[,3]+
beta.vect[4]*X.mat[,4]

# Apply inverse logistic transformation
mu.current<-exp(lp.current)/(1+exp(lp.current))

# Derive variance function and diagonal elements for weight matrix (using
squared
# first differential of link function)
var.i<-(mu.current*(1-mu.current))
g2.i<-(1/(mu.current*(1-mu.current)))^2
W.mat<-diag(1/(var.i*g2.i))

#Calculate information matrix
info.matrix<-t(X.mat)%*%W.mat%*%X.mat

#Derive u terms for score vector
u.i<- (data.DC$CC-mu.current)* (1/(mu.current*(1-mu.current)))

#Calculate score vector

```

```
score.vect<-t(X.mat)%*%W.mat%*%u.i
```

```
#Calculate log likelihood and deviance contribution for current study
```

```
#For convenience, ignore the element of deviance that relates to the full saturated  
# model, because that will cancel out in calculating the change in deviance from  
one
```

```
# iteration to the next (Dev.total - Dev.old [see below]) because the element  
relating
```

```
# to the saturated model will be the same at every iteration).
```

```
log.L<-sum(data.DC$CC*log(mu.current) + (1-data.DC$CC)*log(1-mu.current))
```

```
dev<- -2*log.L
```

```
#Create study specific versions of all key model components
```

```
info.matrix.6<-info.matrix
```

```
score.vect.6<-score.vect
```

```
dev.6<-dev
```

```
nsubs.6<-nsubs
```

```
#Send all of the key model components from the current study to the AC
```

```
save(info.matrix.6,file=paste(AC.Directory,"info.matrix.6.RData",sep=""))
```

```
save(score.vect.6,file=paste(AC.Directory,"score.vect.6.RData",sep=""))
```

```
save(dev.6,file=paste(AC.Directory,"dev.6.RData",sep=""))
```

```
save(nsubs.6,file=paste(AC.Directory,"nsubs.6.RData",sep=""))
```

```
#END STUDY 6
```

```
#####
```

```
#####
```

```
#ITERATION ON ALL LOCAL COMPUTERS NOW COMPLETED
```

```
# KEY MODEL ELEMENTS HAVE BEEN TRANSMITTED TO AC
```

```
#AC WILL NOW USE THESE ELEMENTS TO GENERATE UPDATE TERMS
```

```
#AND TEST FOR CONVERGENCE
```

**#Read back into R, the key elements generated by the local data computers and
#sent to the AC**

```
load(file=paste(AC.Directory,"info.matrix.1.RData",sep=""))  
load(file=paste(AC.Directory,"info.matrix.2.RData",sep=""))  
load(file=paste(AC.Directory,"info.matrix.3.RData",sep=""))  
load(file=paste(AC.Directory,"info.matrix.4.RData",sep=""))  
load(file=paste(AC.Directory,"info.matrix.5.RData",sep=""))  
load(file=paste(AC.Directory,"info.matrix.6.RData",sep=""))
```

```
load(file=paste(AC.Directory,"score.vect.1.RData",sep=""))  
load(file=paste(AC.Directory,"score.vect.2.RData",sep=""))  
load(file=paste(AC.Directory,"score.vect.3.RData",sep=""))  
load(file=paste(AC.Directory,"score.vect.4.RData",sep=""))  
load(file=paste(AC.Directory,"score.vect.5.RData",sep=""))  
load(file=paste(AC.Directory,"score.vect.6.RData",sep=""))
```

```
load(file=paste(AC.Directory,"dev.1.RData",sep=""))  
load(file=paste(AC.Directory,"dev.2.RData",sep=""))  
load(file=paste(AC.Directory,"dev.3.RData",sep=""))  
load(file=paste(AC.Directory,"dev.4.RData",sep=""))  
load(file=paste(AC.Directory,"dev.5.RData",sep=""))  
load(file=paste(AC.Directory,"dev.6.RData",sep=""))
```

```
load(file=paste(AC.Directory,"nsubs.1.RData",sep=""))  
load(file=paste(AC.Directory,"nsubs.2.RData",sep=""))  
load(file=paste(AC.Directory,"nsubs.3.RData",sep=""))  
load(file=paste(AC.Directory,"nsubs.4.RData",sep=""))  
load(file=paste(AC.Directory,"nsubs.5.RData",sep=""))  
load(file=paste(AC.Directory,"nsubs.6.RData",sep=""))
```

#Read in the current beta vector

```
load(file=paste(AC.Directory,"beta.vect.next.RData",sep=""))
```

#Sum the key elements across all studies

```
info.matrix.total<-info.matrix.1+info.matrix.2+info.matrix.3+
  info.matrix.4+info.matrix.5+info.matrix.6
```

```
score.vect.total<-score.vect.1+score.vect.2+score.vect.3+
  score.vect.4+score.vect.5+score.vect.6
```

```
dev.total<-dev.1+dev.2+dev.3+dev.4+dev.5+dev.6
```

```
nsubs.total<-nsubs.1+nsubs.2+nsubs.3+nsubs.4+nsubs.5+nsubs.6
```

```
#Create variance covariance matrix as inverse of information matrix
 #(solve() denotes matrix inversion in R )
```

```
variance.covariance.matrix.total<-solve(info.matrix.total)
```

```
#Create beta vector update terms
```

```
beta.update.vect<-variance.covariance.matrix.total %*% score.vect.total
```

```
#Add update terms to current beta vector to obtain new beta vector for next iteration
```

```
beta.vect.next<-beta.vect.next+beta.update.vect
```

```
#Calculate value of convergence statistic and test whether meets convergence criterion
```

```
converge.value<-abs(dev.total-dev.old)/(abs(dev.total)+0.1)
```

```
if(converge.value<=epsilon)converge.state<-"MET"
```

```
if(converge.value>epsilon)dev.old<-dev.total
```

```
#Now summarise model state after current iteration
```

```
cat("\nSUMMARY OF MODEL STATE after iteration No",iteration.count,
```

```
  "\n\nCurrent deviance",dev.total,"on",
```

```
  (nsubs.total-length(beta.vect.next)), "degrees of freedom",
```

```
  "\nConvergence criterion  ",converge.state,"\n\n")
```

```

cat("Information matrix overall\n")
print(info.matrix.total)

cat("Score vector overall\n")
print(score.vect.total)

#If convergence has been obtained, declare final (maximum likelihood) beta vector,
#and calculate the corresponding standard errors, z scores and p values
 #(the latter two to be consistent with the output of a standard GLM analysis)
#Then print out final model summary

if(converge.value<=epsilon)
{
beta.vect.final<-beta.vect.next
se.vect.final<-sqrt(diag(variance.covariance.matrix.total))
z.vect.final<-beta.vect.final/se.vect.final
pval.vect.final<-2*pnorm(-abs(z.vect.final))

model.parameters<-cbind(beta.vect.final,se.vect.final,z.vect.final,pval.vect.final)
dimnames(model.parameters)<-
list(c("Intercept","BMI","SNP","BMI.456"),c("Coefficient","SE","z-value","p-value"))

model.parameters<-signif(model.parameters,digits=4)

#Print out final model summary
cat("\n\nFINAL MODEL\n")

print(model.parameters)

cat("\nCurrent deviance",dev.total,"on",(nsubs.total-length(beta.vect.next)), "degrees of
freedom","\nAfter iteration No",iteration.count,"\n")
}

```

#Repeat summary of final model state

```
cat("\nSUMMARY OF MODEL STATE after iteration No",iteration.count,
    "\n\nCurrent deviance",dev.total,"on",
    (nsubs.total-length(beta.vect.next)), "degrees of freedom",
    "\nConvergence criterion  ",converge.state,"\n\n")
```

**#Update the stored value of the beta vector to reflect the current estimate - to set
#up the next iteration**

```
save(beta.vect.next,file=paste(AC.Directory,"beta.vect.next.RData",sep=""))
```

C.2.4 Output from Analysis 2:1st Iteration:

$$\hat{b}_{r=1} = [0 \ 0 \ 0 \ 0]$$

All data computers use this coefficient vector for iteration 1

Data Computer 1:

$$I(\hat{b}_{j=1,r=1}) = \begin{bmatrix} 500 & -11.61089 & 0 & 294.75 \\ -11.61089 & 7972.37088 & 0 & -25.55092 \\ 0 & 0 & 0 & 0 \\ 294.75 & -25.55092 & 0 & 387.75 \end{bmatrix}$$

$$S(\hat{b}_{j=1,r=1}) = \begin{bmatrix} -38 & 203.1316 & 0 & 87.5 \end{bmatrix}$$

$$D_{j=1,r=1} = 2772.589$$

Data Computer 2:

$$I(\hat{b}_{j=2,r=1}) = \begin{bmatrix} 750 & -8.417491 & 0 & 443 \\ -8.417491 & 12492.689094 & 0 & -11.777043 \\ 0 & 0 & 0 & 0 \\ -14 & 370.8722 & 0 & 162 \end{bmatrix}$$

$$S(\hat{b}_{j=2,r=1}) = \begin{bmatrix} 443 & -11.777043 & 0 & 578.5 \end{bmatrix}$$

$$D_{j=2,r=1} = 4158.883$$

Data Computer 3:

$$I(\hat{b}_{j=3,r=1}) = \begin{bmatrix} 375 & 34.88511 & 0 & 226.75 \\ 34.88511 & 6407.52995 & 0 & -26.82820 \\ 0 & 0 & 0 & 0 \\ 226.75 & -26.82820 & 0 & 293.75 \end{bmatrix}$$

$$S(\hat{b}_{j=3,r=1}) = \begin{bmatrix} 11 & -14.2244 & 0 & 70.5 \end{bmatrix}$$

$$D_{j=3,r=1} = 2079.442$$

Data Computer 4:

$$I(\hat{b}_{j=4,r=1}) = \begin{bmatrix} 75 & 16.13902 & 16.13902 & 47 \\ 16.13902 & 1265.49746 & 1265.49746 & -12.16424 \\ 16.13902 & 1265.49746 & 1265.49746 & -12.16424 \\ 47 & -12.16424 & -12.16424 & 61.5 \end{bmatrix}$$

$$S(\hat{b}_{j=4,r=1}) = \begin{bmatrix} -8 & 68.06208 & 68.06208 & 12 \end{bmatrix}$$

$$D_{j=4,r=1} = 415.8883$$

Data Computer 5:

$$I(\hat{b}_{j=5,r=1}) = \begin{bmatrix} 500 & 70.56657 & 70.56657 & 297 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 297 & 65.39412 & 65.39412 & 382 \end{bmatrix}$$

$$S(\hat{b}_{j=5,r=1}) = \begin{bmatrix} 36 & 487.2951 & 487.2951 & 149 \end{bmatrix}$$

$$D_{j=5,r=1} = 2772.589$$

Data Computer 6:

$$I(\hat{b}_{j=6,r=1}) = \begin{bmatrix} 175 & 11.5221 & 11.5221 & 102.25 \\ 11.5221 & 2864.847 & 2864.847 & -28.50817 \\ 11.5221 & 2864.847 & 2864.847 & -28.50817 \\ 102.25 & -28.50817 & -28.50817 & 132.25 \end{bmatrix}$$

$$S(\hat{b}_{j=6,r=1}) = \begin{bmatrix} 10 & 149.3701 & 149.3701 & 47.5 \end{bmatrix}$$

$$D_{j=6,r=1} = 970.406$$

Information matrices and score vectors generated by each study are transmitted to AC.

Central Summation at AC:

$$\sum_{j=1}^6 I(\hat{b}_{j,r=1}) = \begin{bmatrix} 2375 & 113.08442 & 98.22769 & 1410.75 \\ 113.08442 & 38649.22602 & 11776.63610 & 39.43446 \\ 98.22769 & 11776.63610 & 11776.63610 & 24.72170 \\ 1410.75 & -39.43446 & 24.72170 & 1835.75 \end{bmatrix}$$

$$\sum_{j=1}^6 S(\hat{b}_{j,r=1}) = \begin{bmatrix} -3 & 1264.5067 & 704.7273 & 528.5 \end{bmatrix}$$

$$\sum_{j=1}^6 D_{j,r=1} = 13169.80$$

Convergence criterion tested: **Not met.**

Derive update vector:

$$I(\hat{\mathbf{b}}_{r=1})^{-1} \mathbf{s}(\hat{\mathbf{b}}_{r=1}) = \begin{bmatrix} -0.32183281 & 0.02228647 & 0.03911561 & 0.53516954 \end{bmatrix}$$

Add update vector to original coefficient vector to produce coefficient vector for second iteration:

$$\hat{\mathbf{b}}_{r=2} = \hat{\mathbf{b}}_{r=1} + I(\hat{\mathbf{b}}_{r=1})^{-1} \mathbf{s}(\hat{\mathbf{b}}_{r=1}) = \begin{bmatrix} -0.32183281 & 0.02228647 & 0.03911561 & 0.53516954 \end{bmatrix}$$

2nd Iteration:

$$\hat{\mathbf{b}}_{r=2} = \begin{bmatrix} -0.32183281 & 0.02228647 & 0.03911561 & 0.53516954 \end{bmatrix}$$

Procedure used in iteration 1 repeated, all data computers use this coefficient vector for iteration 2

For clarity, the information matrix, score vector, and deviance contributions from the individual data computers are omitted from the presentation of this iteration.

Information matrices and score vectors are generated by each study and are transmitted to AC

Central Summation at AC:

$$\sum_{j=1}^6 I(\hat{\mathbf{b}}_{j,r=2}) = \begin{bmatrix} 2295.0536 & 115.0395 & 92.74410 & 1338.4 \\ 115.0395 & 37006.6888 & 11006.04639 & -160.8173 \\ 92.7441 & 11006.0464 & 11006.04639 & -48.61056 \\ 1338.4 & -160.8173 & -48.61056 & 1707.81381 \end{bmatrix}$$

$$\sum_{j=1}^6 S(\hat{\mathbf{b}}_{j,r=2}) = \begin{bmatrix} 4.679958 & 46.098158 & 29.761157 & 17.657043 \end{bmatrix}$$

$$\sum_{j=1}^6 D_{j,r=2} = 12825.07$$

Convergence Criterion: **Not met.**

Derive update vector:

$$I(\hat{\mathbf{b}}_{r=2})^{-1} \mathbf{s}(\hat{\mathbf{b}}_{r=2}) = \begin{bmatrix} -0.0077096093 & 0.0007061835 & 0.0021357681 & 0.0165082231 \end{bmatrix}$$

Add update vector to original coefficient vector to produce coefficient vector for third iteration:

$$\hat{\mathbf{b}}_{r=3} = \hat{\mathbf{b}}_{r=2} + I(\hat{\mathbf{b}}_{r=2})^{-1} \mathbf{s}(\hat{\mathbf{b}}_{r=2}) = \begin{bmatrix} -0.32954242 & 0.02299265 & 0.04125137 & 0.55167776 \end{bmatrix}$$

3rd Iteration:

$$\hat{b}_{r=3} = \begin{bmatrix} -0.32954242 & 0.02299265 & 0.04125137 & 0.55167776 \end{bmatrix}$$

Procedure used in iteration 1 repeated, all data computers use this coefficient vector for iteration 3

For clarity, the information matrix, score vector, and deviance contributions from the individual data computers are again omitted from the presentation of this iteration.

Information matrices and score vectors are generated by each study and are transmitted to AC

Central Summation at AC:

$$\sum_{j=1}^6 I(\hat{b}_{j,r=3}) = \begin{bmatrix} 2290.07166 & 114.04663 & 91.77508 & 1333.57918 \\ 114.04663 & 36898.8956 & 10949.7004 & -169.08819 \\ 91.77508 & 10949.7004 & 10949.7004 & -53.88901 \\ 1333.57918 & -169.0882 & -53.88901 & 1699.55867 \end{bmatrix}$$

$$\sum_{j=1}^6 S(\hat{b}_{j,r=3}) = \begin{bmatrix} 0.02188718 & 0.16208605 & 0.11946433 & 0.05791435 \end{bmatrix}$$

$$\sum_{j=1}^6 D_{j,r=3} = 12824.72$$

Convergence Criterion: **Not met.**

Derive update vector:

$$I(\hat{\mathbf{b}}_{r=3})^{-1} \mathbf{s}(\hat{\mathbf{b}}_{r=3}) = \begin{bmatrix} -2.089082e-10 & 1.660537e-11 & 1.390907e-10 & 5.496857e-10 \end{bmatrix}$$

Add update vector to original coefficient vector to produce coefficient vector for third iteration:

$$\hat{\mathbf{b}}_{r=4} = \hat{\mathbf{b}}_{r=3} + I(\hat{\mathbf{b}}_{r=3})^{-1} \mathbf{s}(\hat{\mathbf{b}}_{r=3}) = \begin{bmatrix} -0.32956275 & 0.02299454 & 0.04126082 & 0.55172828 \end{bmatrix}$$

4th Iteration:

$$\hat{\mathbf{b}}_{r=4} = \begin{bmatrix} -0.32956275 & 0.02299454 & 0.04126082 & 0.55172828 \end{bmatrix}$$

Procedure used in iteration 1 repeated, all data computers use this coefficient vector for iteration 3

For clarity, the information matrix, score vector, and deviance contributions from the individual data computers are omitted from the presentation of this iteration.

Information matrices and score vectors are generated by each study and are transmitted to AC

Central Summation at AC:

$$\sum_{j=1}^6 I(\hat{b}_{j,r=4}) = \begin{bmatrix} 2290.05551 & 114.04125 & 91.77065 & 1333.56304 \\ 114.04125 & 36898.5281 & 10949.48836 & -169.11693 \\ 91.77065 & 10949.48836 & 10949.48836 & -53.90879 \\ 1333.56304 & -169.11693 & -53.90879 & 1699.53140 \end{bmatrix}$$

$$\sum_{j=1}^6 S(\hat{b}_{j,r=4}) = \begin{bmatrix} 2.692875e-07 & 2.018901e-06 & 1.655988e-06 & 6.453095e-07 \end{bmatrix}$$

$$\sum_{j=1}^6 D_{j,r=4} = 12824.72$$

Convergence Criterion: **Met.**

Variance-covariance matrix now obtained by taking the inverse of the summed information matrix

$$\left[\sum_{j=1}^6 I(\hat{b}_{j,r=3}) \right]^{-1} = \begin{bmatrix} 8.054620e-04 & -3.499749e-06 & -6.365439e-06 & -6.325681e-04 \\ -3.499749e-06 & 3.856388e-05 & -3.850815e-05 & 5.362074e-06 \\ -6.365439e-06 & -3.850815e-05 & 1.299160e-04 & 5.283779e-06 \\ -6.325681e-04 & 5.362074e-06 & 5.283779e-06 & 1.085453e-03 \end{bmatrix}$$

and standard errors are obtained by taking the square-root of the diagonal elements of this matrix.

C.2.5 Final Results for Analysis 2:

Coefficient	Estimate	Std Error
Intercept	-0.32960	0.02838
BMI	0.02300	0.00621
BMI.456	0.04126	0.01140
SNP	0.55170	0.03295

Residual deviance: 12824.7 on 9496 degrees of freedom

Bibliography

Aitkin, M., D. Anderson, et al. (1989). Statistical Modelling in GLIM. Oxford, Clarendon Press.

Altman, D. G. and P. Royston (2006). "The cost of dichotomising continuous variables." BMJ **332**(7549): 1080.

Arnett, S. Claas, et al. (2009). "Has pharmacogenetics brought us closer to 'personalized medicine' for initial drug treatment of hypertension?" Current Opinion in Cardiology **24**(4): 333-339.

Arnett, D. K., B. R. Davis, et al. (2005). "Pharmacogenetic Association of the Angiotensin-Converting Enzyme Insertion/Deletion Polymorphism on Blood Pressure and Cardiovascular Risk in Relation to Antihypertensive Treatment: The Genetics of Hypertension-Associated Treatment (GenHAT) Study." Circulation **111**(25): 3374-3383.

Balding, D. J. (2003). "Likelihood-based inference for genetic correlation coefficients." Theoretical Population Biology **63**(3): 221-230.

Beilin, L. J. (1997). "Stress, coping, lifestyle and hypertension: a paradigm for research, prevention and non-pharmacological management of hypertension." Clinical and Experimental Hypertension **19**(5-6): 739-752.

Borenstein, M., L. V. Hedges, et al. (2009). Introduction to meta-analysis. Chichester, John Wiley & Sons.

Brand, E., J. G. Wang, et al. (2003). "An epidemiological study of blood pressure and metabolic phenotypes in relation to the Gbeta3 C825T polymorphism." Journal of Hypertension **21**(4): 729-37.

Braun, R., W. Rowe, et al. (2009). "Needles in the haystack: identifying individuals present in pooled genomic data." PLoS Genetics **5**(10).

Burton, P. and The Wellcome Trust Case Ccontrol Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

Burton, P., I. Fortier, et al. (2010). Biobanks and biobank harmonisation. An Introduction to Genetic Epidemiology. G. Davey-Smith, P. Burton and L. J. Palmer. Bristol, Policy Press.

Burton, P., L. Gurrin, et al. (1998). "Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling." Statistics in Medicine **17**(11): 1261-1291.

Burton, P. R., A. L. Hansell, et al. (2009). "Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology." Int. J. Epidemiol. **38**(1): 263-273.

Buuren, S. v., H. C. Boshuizen, et al. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis." Statistics in Medicine **18**(6): 681-694.

Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association." The Lancet **361**(9357): 598-604.

Cavalli-Sforza, L., P. Menozzi, et al. (1994). The history and geography of human genes. Princeton, Princeton University Press.

Chobanian, A. V., G. L. Bakris, et al. (2003). "Seventh report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure." Hypertension **42**(6): 1206.

Clayton, D. (2010). "On inferring presence of an individual in a mixture: a Bayesian approach." Biostatistics **11**(4): 661-673.

Clayton, D. and M. Hills (1993). Statistical models in epidemiology, Oxford University Press.

Cook, N. R. (1997). "An imputation method for non-ignorable missing data in studies of blood pressure." Statistics in Medicine **16**(23): 2713-2728.

Cook, N. R. (2006). "Imputation strategies for blood pressure data nonignorably missing due to medication use." Clinical Trials **3**(5): 411-420.

Couzin, J. (2008). "GENETIC PRIVACY: Whole-Genome Data Not Anonymous, Challenging Assumptions." Science **321**(5894): 1278-.

Cui, J. and S. Harrap (2002). "Genes and family environment explain correlations between blood pressure and body mass index." Hypertension **40**(1): 7-12.

Cui, J. and S. Harrap (2003). "Antihypertensive treatments obscure familial contributions to blood pressure variation." Hypertension **41**(2): 207-210.

DerSimonian, R. and N. Laird (1986). "Meta-analysis in clinical trials." Controlled Clinical Trials **7**(3): 177-188.

Dobson, A. J. (2002). An introduction to generalized linear models. London, Chapman & Hall/CRC.

Efron, B., T. Hastie, et al. (2004). "Least angle regression." Annals of statistics: 407-451.

Egeland, T., A. E. Fonnelop, et al. (2010). "Complex mixtures: a critical examination of a paper by Homer et al." Forensic Science International: Genetics **in press**.

Eisenstein, B., C. Kauffman, et al. (2009). "Grinding to a Halt: The Effects of the Increasing Regulatory Burden on Research and Quality Improvement Efforts." Clinical Infectious Diseases **49**(3): 328.

ESRC_Secure_Data_Service. (2009). "<http://securedata.ukda.ac.uk/>." Retrieved 12/1/11.

Fisher, R. A. (1966). *The design of experiments*. Edinburgh, Hafner.

Fortier, I., P. R. Burton, et al. (2010). "Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies." International Journal of Epidemiology.

Foster, M. W. and R. R. Sharp (2007). "Share and share alike: deciding how to distribute the scientific and social benefits of genomic data." Nat Rev Genet **8**(8): 633-639.

Gelber, R. D. and A. Goldhirsch (1987). "Interpretation of results from subset analyses within overviews of randomized clinical trials." Statistics in Medicine **6**(3): 371-378.

Gibbs, R. A. and The International HapMap Consortium. (2003). "The International HapMap Project." Nature **426**(6968): 789-796.

Golding, J. (1996). "Children of the Nineties: a resource for assessing the magnitude of long-term effects of prenatal, perinatal and subsequent events." Contemporary Reviews in Obstetrics and Gynaecology **8**: 89-92.

Greely, H. T. (2007). "The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks." Annual Review of Genomics and Human Genetics **8**(1): 343-364.

Halekoh, U., S. Højsgaard, et al. (2006). "The R package geePack for generalized estimating equations." Journal of Statistical Software **15**(2): 1–11.

Hardin, J. W. and J. M. Hilbe (2007). Generalized Estimating Equations, John Wiley & Sons, Inc.

Havlik, R. J., R. J. Garrison, et al. (1979). "Blood pressure aggregation in families." American journal of epidemiology **110**(3): 304.

Hayashi, F. (2000). Econometrics. Princeton ; Oxford, Princeton University Press.

Heath, S. C., I. G. Gut, et al. (2008). "Investigation of the fine structure of European populations with applications to disease association studies." Eur J Hum Genet **16**(12): 1413-1429.

Hedges, L. V. and I. Olkin (1985). Statistical methods for meta-analysis, Academic Press New York.

Higgins, J. P. T., S. G. Thompson, et al. (2003). "Measuring inconsistency in meta-analyses." BMJ **327**(7414): 557-560.

Higgins, J. P. T., A. Whitehead, et al. (2001). "Meta-analysis of continuous outcome data from individual patients." Statistics in Medicine **20**(15): 2219-2241.

Homer, N., S. Szlinger, et al. (2008). "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays." PLoS Genet **4**(8): e1000167.

Hsueh, W. C. Hsueh, et al. (2000). "QTL influencing blood pressure maps to the region of PPH1 on chromosome 2q31-34 in Old Order Amish." Circulation **101**(24): 2810-6.

Hunt, S., L. Atwood, et al. (2002). "Genome scans for blood pressure and hypertension - The National Heart, Lung, and Blood Institute Family Heart Study." Hypertension **40**(1): 1-6.

International HapMap, C. (2003). "The international HapMap project." Nature **426**: 789-796.

Ioannidis, J. P. A., N. A. Patsopoulos, et al. (2007). "Heterogeneity in meta-analyses of genome-wide association investigations." PLoS One **2**(9): 841.

Jacobs, K. B., M. Yeager, et al. (2009). "A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies." Nature Genetics **41**(11): 1253-1257.

Karr, A. F., W. J. Fulp, et al. (2007). "Secure, privacy-preserving analysis of distributed databases." Technometrics **49**(3): 335-345.

Kaye, J. (2005). "Do we need a uniform regulatory system for biobanks across Europe?" Eur J Hum Genet **14**(2): 245-248.

Kaye, J., C. Heeney, et al. (2009). "Data sharing in genomics [mdash] re-shaping scientific practice." Nat Rev Genet **10**(5): 331-335.

Kearney, P. M., M. Whelton, et al. (2005). "Global burden of hypertension: analysis of worldwide data." The Lancet **365**(9455): 217-223.

Koenker, R. (2008). "quantreg: Quantile Regression." **R package version 4.17.** <http://www.r-project.org> (2008).

Last, J. M. (2001). A Dictionary of Epidemiology. Oxford, Oxford University Press.

Law, M. R., N. J. Wald, et al. (2003). "Value of low dose combination treatment with blood pressure lowering drugs: analysis of 354 randomised trials." BMJ **326**(7404): 1427.

Levy, D., A. L. DeStefano, et al. (2000). "Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study." Hypertension **36**(4): 477-83.

Levy, D., G. B. Ehret, et al. (2009). "Genome-wide association study of blood pressure and hypertension." Nat Genet **41**(6): 677-687.

Lewington, S., R. Clarke, et al. (2002). "Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies." Lancet **360**(9349): 1903-13.

Liang, K.-Y. and S. L. Zeger (1986). "Longitudinal data analysis using generalized linear models." Biometrika **73**(1): 13-22.

Lifton, R. P., A. G. Gharavi, et al. (2001). "Molecular Mechanisms of Human Hypertension." Cell **104**(4): 545-556.

Little, J., J. P. T. Higgins, et al. (2009). "STrengthening the REporting of Genetic Association Studies (STREGA)-An Extension of the STROBE Statement." Genetic Epidemiology **33**(7): 581-598.

Lowrance, W. W. and F. S. Collins (2007). "Identifiability in genomic research." Science **317**(5838): 600.

Lunshof, J. E., R. Chadwick, et al. (2008). "From genetic privacy to open consent." Nature Reviews Genetics **9**(5): 406-410.

Malfroy, M., C. A. Llewelyn, et al. (2004). "Using patient-identifiable data for epidemiological research." Transfusion Medicine **14**(4): 275-279.

Marchini, J., L. R. Cardon, et al. (2004). "The effects of human population structure on large genetic association studies." Nature Genetics **36**(5): 512-517.

Marchini, J., C. Spencer, et al. (2007). A Bayesian hierarchical mixture model for genotype calling in a multi-cohort study, (in preparation).

Martin, N. (2000). Gene-environment interaction and twin studies. London, Greenwich Medical Media Ltd

Masca, N., N. A. Sheehan, et al. (2011). "Pharmacogenetic interactions and their potential effects on genetic analyses of blood pressure." Statistics in Medicine **30**(7): 769-783.

McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature Reviews Genetics **9**(5): 356-369.

McClelland, R. L., R. A. Kronmal, et al. (2008). "Estimation of risk factor associations when the response is influenced by medication use: An imputation approach." Statistics in Medicine **27**(24): 5039-5053.

McCullagh, P. J. and J. A. Nelder (1991). Generalized linear models, Chapman & Hall.

Murray, C. J. L. and A. D. Lopez (1997). "Mortality by cause for eight regions of the world: Global Burden of Disease Study." The Lancet **349**(9061): 1269-1276.

Narula, S. C., P. H. Saldiva, et al. (1999). "The minimum sum of absolute errors regression: a robust alternative to the least squares regression." Statistics in Medicine **18**(11): 1401-17.

Negassa, A. and J. A. Hanley (2007). "The effect of omitted covariates on confidence interval and study power in binary outcome analysis: A simulation study." Contemporary Clinical Trials **28**(3): 242-248.

Newton-Cheh, C., T. Johnson, et al. (2009). "Genome-wide association study identifies eight loci associated with blood pressure." Nat Genet **41**(6): 666-676.

Olkin, I. and A. Sampson (1998). "Comparison of Meta-Analysis Versus Analysis of Variance of Individual Patient Data." Biometrics **54**(1): 317-322.

P3G_Consortium, G. Church, et al. (2009). "Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection." PLoS Genet **5**(10): e1000665.

Paik, M. C. (1988). "Repeated measurement analysis for nonnormal data in small samples." Communications in Statistics-Simulation and Computation **17**(4): 1155-1171.

Petrie, J. C., E. T. O'Brien, et al. (1986). "Recommendations on blood pressure measurement." British Medical Journal (Clinical research ed.) **293**(6547): 611.

Pickering, T. G. (1997). "The effects of environmental and lifestyle factors on blood pressure and the intermediary role of the sympathetic nervous system." Journal of human hypertension **11**: S9.

Power, C. and J. Elliott (2006). "Cohort profile: 1958 British birth cohort (National Child Development Study)." International Journal of Epidemiology **35**(1): 34.

Pritchard, J. K. and P. Donnelly (2001). "Case-control studies of association in structured or admixed populations." Theoretical Population Biology **60**(3): 227-237.

Purcell, S. M., N. R. Wray, et al. (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." Nature.

Quinn, G. P. and M. J. Keough (2002). Experimental design and data analysis for biologists. Cambridge, Cambridge University Press.

Raskin, P. (2003). "Treatment of Hypertension in Adults With Diabetes." Clinical Diabetes **21**(3): 120-121.

Raum, E., S. Lietzau, et al. (2008). "For the majority of patients with diabetes blood pressure and lipid management is not in line with recommendations. Results from a large population-based cohort in Germany." Pharmacoepidemiology and Drug Safety **17**(5): 485-494.

Reiner, A. P., E. Ziv, et al. (2005). "Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study." The American Journal of Human Genetics **76**(3): 463-477.

Resnik, D. B. (2010). "Genomic research data: open vs. restricted access." IRB **32**(1): 1.

Rice, T. Rice, et al. (2000). "Genome-wide linkage analysis of systolic and diastolic blood pressure: the Quebec Family Study." Circulation **102**(16): 1956-63.

Riley, R. D., P. C. Lambert, et al. (2010). "Meta-analysis of individual participant data: rationale, conduct, and reporting." BMJ **340**.

Royall, R. M. (1986). "Model robust confidence intervals using maximum likelihood estimators." International Statistical Review/Revue Internationale de Statistique **54**(2): 221-226.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, J. Wiley & Sons.

Sampson, J. and H. Zhao (2009). "Identifying Individuals in a Complex Mixture of DNA with Unknown Ancestry." Statistical Applications in Genetics and Molecular Biology **8**(1): 37.

Sankararaman, S., G. Obozinski, et al. (2009). "Genomic privacy and limits of individual detection in a pool." Nature Genetics **41**(9): 965-967.

Smith, V. (2009). "Data publication: towards a database of everything." BMC Research Notes **2**(1): 113.

Stahl, E. A., S. Raychaudhuri, et al. (2010). "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci." Nat Genet **42**(6): 508-514.

Sutton, A. J., K. R. Abrams, et al. (2000). Methods for meta-analysis in medical research, Wiley, London.

Sutton, A. J., D. Kendrick, et al. (2008). "Meta analysis of individual and aggregate level data." Statistics in Medicine **27**(5): 651-669.

Thorisson, G. A., J. Muilu, et al. (2009). "Genotype-phenotype databases: challenges and solutions for the post-genomic era."

Tobin, M. D., N. A. Sheehan, et al. (2005). "Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure." Statistics in Medicine **24**(19): 2911-35.

Trinidad, S. B., S. M. Fullerton, et al. (2010). "Genomic research and wide data sharing: views of prospective participants." Genetics in medicine: official journal of the American College of Medical Genetics **12**(8): 486.

Turner, G. L. Schwartz, et al. (2001). "Antihypertensive pharmacogenetics: Getting the right drug into the right patient." Journal of Hypertension **19**(1): 1-11.

Venables, W. N. and B. D. Ripley (2002). Modern Applied Statistics with S. New York, Springer.

Visscher, P. M., W. G. Hill, et al. (2009). "The limits of individual identification from sample allele frequencies: theory and statistical analysis." PLoS Genet **5**(9): e1000628.

Vora, P. Ouyang, et al. (2008). "Racial differences of lipoprotein subclass distributions in postmenopausal women." Ethnicity **18**(2): 176-80.

Wallace, S., S. Lazor, et al. (2009). "Consent and population genomics: the creation of generic tools." IRB: Ethics & Human Research **31**: 15-20.

Wang, J., J. Kuusisto, et al. (2007a). "Variants of transcription factor 7-like 2 (TCF7L2) gene predict conversion to type 2 diabetes in the Finnish Diabetes Prevention Study and are associated with impaired glucose regulation and impaired insulin secretion." Diabetologia **50**(6): 1192-1200.

Wang, Y. R., G. C. Alexander, et al. (2007b). "Outpatient Hypertension Treatment, Treatment Intensification, and Control in Western Europe and the United States." Archives of Internal Medicine **167**(2): 141.

White, H. (1982). "Maximum likelihood estimation of misspecified models." Econometrica: Journal of the Econometric Society **50**(1): 1-25.

White, I. R., N. Chaturvedi, et al. (1994). "Median analysis of blood pressure for a sample including treated hypertensives." Statistics in Medicine **13**(16): 1635-41.

White, I. R., I. Koupilova, et al. (2003). "The use of regression models for medians when observed outcomes may be modified by interventions." Statistics in Medicine **22**(7): 1083-96.

Whitehead, A., R. Z. Omar, et al. (2001). "Meta analysis of ordinal outcomes using individual patient data." Statistics in Medicine **20**(15): 2243-2260.

Wilcox, J. (1961). "Observer Factors in the Measurement of Blood Pressure." Nursing Research **10**(1): 4-17.

Williams, R. L. (2000). "A Note on Robust Variance Estimation for Cluster-Correlated Data." Biometrics **56**(2): 645-646.

Wolf-Maier, K., R. S. Cooper, et al. (2003). "Hypertension Prevalence and Blood Pressure Levels in 6 European Countries, Canada, and the United States." JAMA **289**(18): 2363-2369.

Wolfson, M., S. E. Wallace, et al. (2010). "DataSHIELD: resolving a conflict in contemporary bioscience - performing a pooled analysis of individual-level data without sharing the data." International Journal of Epidemiology.

Wong, M. Y., N. E. Day, et al. (2003). "The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement?" Int. J. Epidemiol. **32**(1): 51-57.

Wright, S. (1968). Evolution and the Genetics of Populations, University of Chicago Press.

Yan, J. and J. Fine (2004). "Estimating equations for association structures." Statistics in Medicine **23**(6): 859-874.

Yang, Q., F. Sun, et al. (2007). "Maternal influence on blood pressure suggests involvement of mitochondrial DNA in the pathogenesis of hypertension: the Framingham Heart Study." Journal of Hypertension **25**(10): 2067-2073.

Zeger, S. L. and K. Y. Liang (1986). "Longitudinal Data Analysis for Discrete and Continuous Outcomes." Biometrics **42**(1): 121-130.

Zeger, S. L., K. Y. Liang, et al. (1988). "Models for longitudinal data: a generalized estimating equation approach." Biometrics: 1049-1060.

Zeggini, E., L. J. Scott, et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nature Genetics **40**(5): 638-645.

Zink, A. and A. J. Silman (2008). "Ethical and legal constraints on data sharing between countries in multinational epidemiological studies in Europe report from a joint workshop of the European League Against Rheumatism standing committee on epidemiology with the "AutoCure" project." Annals of the Rheumatic Diseases **67**(7): 1041-1043.

Appendum

This section contains the three publications that support this thesis.

The first paper (Masca *et al.*, 2011) was published in *Statistics in Medicine* and concerns some of the work reported in Chapter 1.

The second paper, Masca *et al.* (in press) has been accepted for publication in the *International Journal of Epidemiology* and concerns the work in Chapter 2.

The third paper (Wolfson *et al.*, 2010) outlines the work reported in Chapter 3, and was published in the *International Journal of Epidemiology*.

Pharmacogenetic interactions and their potential effects on genetic analyses of blood pressure

Nicholas Masca,^{*†} Nuala A. Sheehan and Martin D. Tobin

Background: In observational studies, analyses of blood pressure (BP) typically require some correction for the use of antihypertensive medications by study participants. Different approaches to correcting for treatment have been compared, but the impact of pharmacogenetic interactions that influence the efficacy of antihypertensive treatments on estimates of genetic main effects has not been considered. This work demonstrates the potential influence of pharmacogenetic interactions in genetic analyses of BP.

Methods: A simulation study is conducted to test the influence of pharmacogenetic interactions on approaches to the analysis of BP. Results from three plausible scenarios are presented.

Results: Informative BP approaches (*Fixed Treatment Effect, Non-parametric adjustment, Censored Normal Regression*) perform well when there is no pharmacogenetic interaction, but yield biased estimates of the main effects of particular genetic variants when pharmacogenetic interactions exist. Substitution approaches (*Binary Trait, Fixed Substitution, Random Substitution, Median Method*) are unaffected by pharmacogenetic interactions, but consistently perform sub-optimally.

Conclusions: We recommend that the Informative BP approaches remain the most appropriate methods to use in practice, but stress that caution is required in the interpretation of their results—especially when an interaction between treatment and a genetic variant of interest is suspected. We make some suggestions as to how to check for possible interactions and confirm the results from genetic analyses of BP, but warn that these should be reviewed when data on real pharmacogenetic interactions become available. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: blood pressure; genetic association; pharmacogenetics; treatment effects; imputation; bias

1. Introduction

Hypertension (high blood pressure (BP)) is a common condition estimated to affect over 25 per cent of adults worldwide [1]. Although hypertension itself is asymptomatic, it is a major contributor to the risk of cardiovascular disease, which accounts for up to 30 per cent of all deaths [2, 3]. Even changes within the normal range of BP are associated with risk of stroke and coronary heart disease (CHD) [4]. BP in its own right is therefore of major importance to public health.

Lifestyle factors such as dietary salt intake, physical activity, smoking, and body-mass index (BMI) are all known to influence BP [5, 6], but BP also has a substantial heritable component [7, 8]. Identification of the genetic determinants of BP can offer insights into the biological pathways underpinning BP regulation [9], and, indeed, this has been a key aim of recent genetic association studies of BP.

Paramount to the success of a genetic association study is a sufficient statistical power to detect the generally modest effects of common genetic variants [10, 11]. In genome-wide association studies, hundreds of thousands or even millions of genetic variants are tested for association with the phenotype of interest, and an allowance for multiple testing must be made. The threshold for genome-wide

Department of Health Sciences, 2nd Floor, Adrian Building, University of Leicester, University Road, Leicester LE1 7RH, U.K.

*Correspondence to: Nicholas Masca, Department of Health Sciences, 2nd Floor, Adrian Building, University of Leicester, University Road, Leicester LE1 7RH, U.K.

†E-mail: nm180@leicester.ac.uk

significance is currently defined as $p < 5 \times 10^{-8}$ [12]. A sufficiently large sample size is crucial to the provision of an adequate power to detect associations for BP at this threshold. Recent breakthroughs in genome-wide association studies of BP have been achieved using large sample sizes [13, 14]. For instance, the Global BPgen Consortium [13] meta-analysed 17 cohorts consisting of a total of 34 433 participants, and the CHARGE Consortium [14] meta-analysed five cohorts consisting of 29 136 participants. The single nucleotide polymorphisms (SNPs) highlighted in these studies had reported effect sizes between approximately 0.5 and 1 mmHg per copy of the minor allele for systolic BP, and approximately 0.35–0.5 mmHg per copy of the minor allele for diastolic BP (typically about $\frac{1}{40}$ th to $\frac{1}{15}$ th of a standard deviation).

Besides sample size, there are several other factors that may limit the statistical power of genetic association studies of BP. For instance, it can be difficult to gain a reliable measure of an individual's BP [10] because BP varies in different situations and at different time points throughout the day. Other measurement difficulties, such as an alerting (or 'white-coat') response and observer bias (including 'digit preference', which entails rounding BP up or down), can also influence recordings of BP [15, 16]. Moreover, investigations into the aetiology of BP are affected by the use of antihypertensive treatments by study participants. Since hypertension is highly prevalent within western countries, drugs to lower BP—antihypertensives—are widely prescribed. Population-based cohort studies therefore could have up to a quarter of participants on antihypertensive treatment (or even more in studies of older populations) [13, 14]. For these treated participants, any BP measurements provided within a study will reflect 'modified BP' values, as opposed to the 'underlying BP' values that exist, in principle, in the absence of treatment. It has been shown that a failure to adequately correct for the inclusion of *modified BPs* within an analysis can distort the results [17–21]. Because antihypertensive treatments lower BP, *modified BPs* will be lower than the unobserved, *underlying BPs*, and the results from an analysis may thus be misleading if no account is made for antihypertensive use. Previous work has compared different approaches to correct analyses for the use of antihypertensive medications [20]. One approach that was recommended and is particularly easy to use was proposed by Cui *et al.* ('Fixed Treatment Effect'—see Section 2) [22, 23], and was actually employed in both the Global BPgen and CHARGE consortia [13, 14].

In studies of BP, a situation known as a 'differential intervention' occurs when either the threshold for receiving treatment or the effect of treatment depends on another variable [19]. For example, a differential threshold for receiving treatment occurs where antihypertensive medications are prescribed to individuals with diabetes at a lower threshold of BP than to non-diabetics [24]. A differential treatment effect would occur, for example, if the efficacy of treatment depends on genotype for a particular genetic variant (a type of *pharmacogenetic* interaction) [25, 26].

White *et al.* [19] investigated how three different approaches to correct for the use of treatment ('No Adjustment', 'Exclude' and 'Median Method'—see Section 2) perform when there is a differential intervention. They looked at both a differential threshold for receiving treatment and a differential treatment effect and found that, in each situation, the approaches tested behaved very differently to one another. Similarly, a more recent study, which tested a number of additional approaches under a differential threshold for receiving treatment, found that several approaches led to biased estimates of main effects of interest [21]. It thus appears that a differential intervention can affect the performance of different approaches to correct for the use of treatment.

We aim to investigate the plausible scenario where a genetic variant leads to a differential treatment effect (rather than a differential threshold for receiving treatment) in a cross-sectional genetic association study of BP. In particular, we aim to assess how the different approaches to correct for the use of treatment perform in such settings, when the primary interest of the analysis is in estimating the marginal effect of a particular genetic variant on BP (i.e. the effect of the variant unmitigated by treatment). This is an important issue; if the approaches to analysis that have been previously recommended [20] produce biased estimates in the presence of a pharmacogenetic interaction involving a differential treatment effect, it would be necessary to reconsider which method to use when such an interaction is suspected.

This paper demonstrates how the existence of pharmacogenetic interactions that influence the efficacy of antihypertensives can exacerbate difficulties in analyses of BP when some participants are on treatment. Because the precise genetic variants that alter the efficacy of antihypertensives are currently unknown, there are, at present, no real data with which to test our findings. Such evidence will eventually be forthcoming, however. In order to explore these issues, we consider a number of plausible hypothetical situations in simulation studies.

In Section 2, current approaches to the analysis of BP are outlined. Our simulation study is then described in Section 3, and results are provided in Section 4. Recommendations are provided in Section 5, along with a discussion of the implications and limitations of our work.

2. Approaches to analysis

'Naïve' approaches to the analysis of BP [21] include: (a) ignoring the problem of antihypertensive use altogether and analysing all the observed BP data without regard to the use of treatment [*No Adjustment*]; (b) adjusting for the use of antihypertensives during analysis by modelling treatment as a binary covariate [*Treatment as a Binary Covariate*]; and (c) omitting any participant who uses antihypertensives from analysis [*Exclude*]. Although these approaches have been widely used in practice (e.g. [27–29]), previous work has shown that they often lead to biased results [18–21].

We classify a second group of approaches to the analysis of BP data as 'Substitution' approaches. These approaches assume that any participants on antihypertensive treatment are hypertensive, and they typically substitute modified BPs (i.e. BP measurements for participants on antihypertensives) for alternative values. The Substitution approaches include: (d) dichotomising the quantitative measures of BP and performing a logistic regression, using a dichotomous hypertension outcome (Yes = participant uses antihypertensive medication or has SBP greater than or equal to some accepted threshold, such as 140 mmHg; No = otherwise) [*Binary Trait*]; (e) substituting modified BPs for a constant, b , which, for example, is set to 140 mmHg—the minimum threshold of SBP for a clinical diagnosis of hypertension [3] [*Fixed Substitution* [30]]; (f) substituting modified BPs for values generated randomly from a distribution consistent with Stage 1 (mild) hypertension (such as a normal distribution with mean 150 mmHg and standard deviation 5, truncated at 140 and 160) [*Random Substitution* [30]]; and (g) substituting modified BPs for some value k , and fitting a quantile regression to the data [31] (*Median Method* [17, 19]). For (g), k is selected to be a value at the upper end of the realistic distribution of BP (such as 160–200 mmHg for SBP), and, crucially, participants who use antihypertensives are assumed to have *underlying BP* above the median.

We classify three remaining approaches as 'Informative BP' approaches. These approaches use all the observed data in the analyses and apply a simple mathematical correction either to modified BPs themselves or to the statistical likelihood function. The Informative BP approaches include: (h) adding a fixed constant, c , to modified BPs, which is set to represent an average (negative) antihypertensive treatment effect (*Fixed Treatment Effect* [22, 23]); (i) adjusting BP measurements by adding the difference between a *raw* and an *adjusted residual* (see below) to each observation (*Non-Parametric Adjustment* [8, 20]); and (j) assuming that modified BPs are right-censored, and fitting a tobit-type model to the data (*Censored Normal Regression*). For (i), an algorithm is applied to each individual BP reading in turn to derive a set of adjusted residuals from the raw residuals. To illustrate this approach simply, we assume that the raw residuals are from a null regression model (including only an intercept term without covariates); such an approach may also be used in practice where one wishes to retain flexibility to adjust for covariates at a later stage of the analysis. For untreated individuals, the adjusted residual is simply equal to the raw residual, but for treated individuals, the adjusted residual is an average of the raw residual for that particular individual and any greater, adjusted residuals. Hence, if an individual is treated the adjusted residual is greater than the raw residual (with the exception of the individual with the highest BP—whose adjusted residual is equal to the raw residual), and the observed SBP is adjusted upward. Approaches (i) and (j) are methodologically related; where (i) *averages* over the probability density function to the right of any modified BPs, (j) *integrates* over this area.

The approaches within the Substitution and Informative BP groups are basically imputation methods, because they impute BP for those measures distorted by treatment. A further group of approaches, however, have specifically been referred to as 'Multiple Imputation' approaches [21, 32]. These Multiple Imputation approaches impute modified BPs by conditioning on out-of-study or pre-treatment measurements [18, 21, 33], which are typically only available in longitudinal studies. Since many studies of BP (for example, in the Global BPgen consortium [13], which is the type of application we have in mind) have little or no longitudinal data, these approaches often cannot be applied. Our focus in this paper is on approaches that *can* be applied in cross-sectional data; that is, our focus is on the approaches within the Naïve, Substitution and Informative BP groups.

Table I. Summary of the analysis models with parameter values used in the simulated studies. Shaded and non-shaded regions denote the three categories of approaches: Naïve, Substitution, and Informative BP.

	Methods	Description	Parameter values
<i>Naïve</i>			
(a)	No Adjustment	Ignore use of treatment; analyse all observations in a linear regression model	
(b)	Treatment as a Binary Covariate	Adjust for antihypertensive treatment use by fitting $TREAT_i$ as a binary covariate	$TREAT_i = 1$ if individual i uses antihypertensive medication; $TREAT_i = 0$ otherwise
(c)	Exclude	Exclude any participants who use antihypertensive medication from the analysis	
<i>Substitution</i>			
(d)	Binary Trait	Define a binary ‘hypertension’ outcome, and fit a logistic regression model to the data	$hypertension_i = 1$ if $TREAT_i = 1$ or if $Z_i \geq 140$ mmHg; $hypertension_i = 0$ otherwise
(e)	Fixed Substitution	Substitute <i>modified BPs</i> for the constant b	$b = 130$ and 140 mmHg
(f)	Random Substitution	Substitute <i>modified BPs</i> for values generated randomly from a normal distribution	$\sim N(150, 5^2)$ truncated to $[140, 160]$
(g)	Median Method	Substitute <i>modified BPs</i> for the value k , and fit a quantile regression to the data	$k = 140, 150, 160$ in Scenario 1; $k = 160, 180$ and 200 in subsequent scenarios
<i>Informative BP</i>			
(h)	Fixed Treatment Effect	Add the constant c to <i>modified BPs</i>	$c = 5; 10; \text{ and } 15$
(i)	Non-parametric Adjustment	Apply an algorithm to derive a set of adjusted residuals; adjust <i>modified BPs</i> by adding the difference between the current adjusted and raw residuals	
(j)	Censored Normal Regression	Assume that <i>modified BPs</i> are right-censored; fit a tobit-model to the data	

Some of the Substitution and Informative BP approaches require specific values to be specified [such as the constant b for approach (e), and the constant c for approach (h)]. Guidance for selecting these imputation values has typically been provided by the original proposers (e.g. in [17, 19, 22, 23, 30]), and depends on knowledge about the specific drugs under consideration. A number of different values can potentially be used with each approach, however. In our simulations we use a range of illustrative parameter values, which we outline in the following section. Table I provides a brief description of each approach and lists the parameter values used for each in our simulations.

3. Simulation study

The aim in our simulation studies is to test how different, realistic scenarios impact upon the approaches to correct for the use of antihypertensives. We are interested in estimating the marginal effect of a particular genetic variant on the underlying BP in the whole study population (i.e. the main effect of a SNP on the BP that would have prevailed in the absence of antihypertensive treatment taken by a proportion of the population). Three scenarios are simulated and Monte Carlo estimates of the statistical power, the type I error rate, and the mean level of bias are derived with respect to a SNP for each approach. A SNP is a genetic variant that has one of two possible alleles on each of the two

homologous chromosomes. For any particular SNP, the allele that has the lowest frequency within a given population is known as the *minor allele*. Thus, for a particular SNP, an individual may have zero, one, or two copies of the minor allele.

In the first scenario, ‘Non-differential intervention’, the approaches are compared in a situation where both the threshold for receiving treatment and the effect of treatment do not depend on any other factor. This is a ‘baseline’ type scenario, which determines the potential levels of performance attainable by each approach. The first scenario forms the basis of the subsequent scenarios, which are created by altering one or more of its properties. A full description of Scenario 1 is thus provided below, while for subsequent scenarios, only those properties that differ from Scenario 1 are described.

3.1. Scenario 1: non-differential intervention

Scenario 1 is designed to represent a population-based study of BP, consisting of 2000 unrelated participants aged between 25 and 80 years. For the i th participant ($i = 1, \dots, 2000$), Z_i denotes underlying SBP (in mmHg); AGE_i denotes age (in years); SEX_i denotes sex (1 = male; 0 = female); $SNP1_i$ denotes genotype for a SNP with minor allele frequency 0.3 (zero, one, or two copies of the minor allele)—which is centred for comparability across different SNP effect sizes; and ε_i denotes normally distributed random error. AGE_i is generated from a uniform distribution with parameters 25 and 80; SEX_i is generated from a Bernoulli distribution with probability 0.5; and $SNP1_i$ is generated from two Bernoulli trials with probability 0.3. For any individual, the underlying SBP Z_i is simulated from a linear regression model with an additive genetic effect (i.e. where two copies of the minor allele yield twice the effect of one copy):

$$Z_i = \beta_0 + \beta_1 AGE_i + \beta_2 SEX_i + \beta_3 SNP1_i + \varepsilon_i \quad (1)$$

where $\beta_0 = 110$ is an intercept coefficient, $\beta_1 = 0.4$, $\beta_2 = 3$, $\beta_3 = +2/-2/0$, and $\varepsilon_i \sim N(0, 18^2)$. Note that the simulation sample size of 2000 individuals is a realistic size for a cohort that would be part of a larger consortium. However, the simulated effect size of SNP1 is larger than would typically be expected, but is required here to ensure that the analysis models are adequately powered.

An individual with underlying SBP greater or equal to 140 mmHg is labelled as hypertensive [3], and will possibly receive treatment. In practice, not all hypertensive individuals receive antihypertensive medication; we therefore assume that antihypertensives are received with probability 0.75. Hence, $TREAT_i$ denotes treated status (1 = Yes; 0 = No), and is generated from a Bernoulli distribution with probability 0.75. As treatment is only administered to hypertensives, $TREAT_i$ is always 0 if $Z_i < 140$.

For individual i , an *observed SBP*, Y_i , is generated to represent the BP measurements typically obtained within studies. For individuals who use antihypertensives, Y_i is derived by subtracting a treatment effect from Z_i . We shall refer to the size of the treatment effect as γ_i , and generate it from a normal distribution with mean 15 mmHg and variance 4^2 [i.e. $\gamma_i \sim N(15, 4^2)$]. To prevent any unrealistic cases, in which the treatment directly *increases* SBP, γ_i is truncated at 0.

We simulate 1000 data sets under the null hypothesis of no SNP1 effect on BP (i.e. when $\beta_3 = 0$), and under two alternative hypotheses (i.e. when $\beta_3 = 2$ and when $\beta_3 = -2$). Monte Carlo estimates of the type I error rate and the statistical power are derived for each approach with respect to the marginal effect of SNP1 on SBP. Mean bias (in mmHg) is also calculated with respect to estimates of the effect of SNP1, but we summarize the results by reporting mean bias when $\beta_3 = 2$ only. Note that the measures of power, type I error, and bias will depend on factors such as the sample size, the minor allele frequency of the SNP of interest, the size of the SNP effect, the proportion of individuals treated within each study, the magnitude of the treatment effect, etc. Descriptive statistics for the studies simulated in this scenario are presented in Table II. Note that the mean proportion of individuals who receive antihypertensives within these studies is relatively high at 27.87 per cent. In order to illustrate the potential implications of the interactions of interest we explore, this was deliberately simulated at a relatively high level. Although there are examples of studies with a similar proportion of individuals on treatment (such as studies of older populations [34]), we accept that this proportion is towards the upper end of the scale.

Note that statistical power is assessed relative to the simulated SNP effect sizes of +2 (i.e. $\beta_3 = 2$) and -2 (i.e. $\beta_3 = -2$), and type I error is assessed relative to the simulated SNP effect of 0 (i.e. $\beta_3 = 0$). As other authors have done [20, 35, 36], we chose parameter coefficients to generate data sets that were realistic representations of many real epidemiological studies of BP (see Table II).

Table II. Descriptive statistics for Scenario 1—non-differential intervention.

Summary statistics	Scenario 1
Sample size	2000
Mean underlying SBP (SD)	133.71 (19.2)
For <i>treated</i> subjects	153.24 (10.4)
For <i>untreated</i> subjects	126.17 (16.2)
Mean observed SBP (SD)	129.53 (16.0)
For <i>treated</i> subjects	138.24 (11.2)
For <i>untreated</i> subjects	126.17 (16.2)
Percentage of individuals with SBP > 140	38.17
Percentage of individuals treated	27.87
Mean treatment effect (SD)	15.00 (4.0)

3.2. Scenario 2: extreme pharmacogenetic interaction

In Scenario 2, a differential treatment effect is simulated, and the efficacy of treatment depends on SNP1. A pharmacogenetic interaction is therefore implemented here. As with results reported by Turner *et al.* [37], the pharmacogenetic interaction *reduces* the efficacy of treatment in the presence of the minor allele.

As in Scenario 1, the underlying SBP, Z_i , is simulated from equation (1), and treatment is allocated to hypertensive individuals with probability 0.75. In contrast to Scenario 1, however, the distribution from which a treatment effect is generated depends on the genotype for SNP1. Hence, the treatment effect, γ_i , is generated from $N(18, 4^2)$, $N(13.43, 4^2)$, or $N(9, 4^2)$ corresponding to whether the i th individual has zero, one, or two copies of the minor allele, respectively. These particular distributions for the treatment effect are chosen such that the overall mean treatment effect remains equal to 15 mmHg. The descriptive statistics generated in this scenario are therefore roughly the same as in Scenario 1 (and thus, are not provided). All treatment effects are again truncated at zero.

3.3. Scenario 3: pharmacogenetic interaction with one class of antihypertensive

In the previous scenario, the simulated pharmacogenetic interaction is assumed to affect all individuals receiving treatment. Given that different classes of antihypertensive medication are commonly used, and that different classes of antihypertensives affect different biological pathways, it is probably unrealistic to assume that a given genetic variant will interact with all treatment types. The influence of the pharmacogenetic interaction represented in Scenario 2 is therefore likely to be more extreme than that of a real pharmacogenetic interaction in a real genetic association study of BP. In Scenario 3, two different classes of treatment are therefore simulated, and only one of these classes of treatment interacts with SNP1.

As before, the underlying SBP, Z_i , is generated from the model in equation (1), and hypertensive individuals are allocated treatment with probability 0.75. In this scenario, however, two classes of treatment are simulated. Treated participants are randomized either to receive Treatment A or Treatment B with the probabilities 0.33 and 0.67, respectively. Treatment A represents angiotensin-converting enzyme (ACE) inhibitors, a common class of antihypertensive medication [36], and Treatment B represents usage of any other antihypertensives (pooled together). For Treatment A, the treatment effect is dependent on SNP1, and a pharmacogenetic interaction is implemented in the same way as that implemented in Scenario 2. Hence, for individuals on Treatment A, γ_i is generated from $N(18, 4^2)$, $N(13.43, 4^2)$, or $N(9, 4^2)$ corresponding to whether the i th individual has zero, one, or two copies of the minor allele, respectively; for individuals on Treatment B, the treatment effect is independent of SNP1, and is thus generated as in Scenario 1 [i.e. $\gamma_i \sim N(15, 4^2)$]. All treatment effects are again truncated at zero.

4. Results

As noted earlier, the focus of our analyses is on the marginal effect of the genetic variant SNP1 on BP. The model in equation (1) is fitted for all approaches [except *Treatment as a Binary Covariate* (b) and *Binary Trait* (d)] by replacing the underlying SBP, Z_i , with the values imputed according to each method (e.g. for *No Adjustment* (a), these would simply be the observed SBPs, Y_i , for all individuals). Since this model is as close as one can get to the true generating model in absence of knowledge

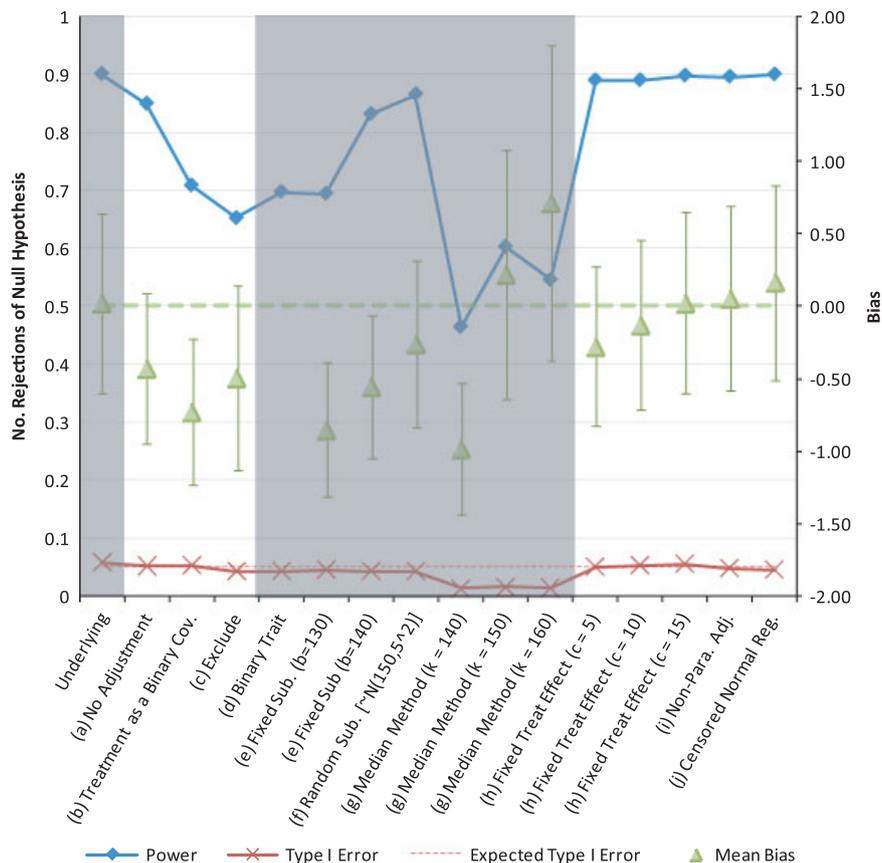


Figure 1. Results of Scenario 1—‘Non-differential Intervention’. Approaches are arranged in categories (from left to right): *Naïve*, *Substitution*, *InformativeBP*, and are highlighted by the shaded regions. The left-most approach is the analysis of underlying SBP, which is included for comparison purposes. Power (with respect to the simulated SNP1 effect of +2) is denoted in blue diamonds; type I error (with respect to the simulated SNP1 effect of 0) is denoted in red crosses. Powers and levels of type I error are evaluated on the left-vertical axis, and lines joining the points are displayed for clarity. Mean bias (with SE) with respect to $\beta_3=2$ is displayed in green triangles, and is evaluated on the right-vertical axis (in mmHg). Because (d) fits a logistic regression and yields log-odds ratios, estimates of the effect of SNP1 are not comparable to the other approaches and are omitted from the plot. Note that power with respect to the simulated SNP1 effect of -2 is not shown for this scenario because it is no different than the power with respect to the simulated SNP1 effect of $+2$.

of a pharmacogenetic interaction, we note that this represents a ‘best-case’ scenario. For method (b), equation (1) is fitted with the observed SBPs, Y_i , as outcome and an additional binary covariate for treatment, $TREAT_i$, is included on the right-hand side. For method (d), a logistic regression model is fitted to the dichotomous outcome, hypertension $_i$ (see Table I).

Results for Scenarios 1–3 are summarized graphically in Figures 1–3, respectively. In each figure and for each approach, the statistical power to detect the marginal effect of SNP1 and the type I error are shown on the left-vertical axis (at the 5 per cent level of significance), and the mean bias of the estimated coefficient of SNP1 (in mmHg), with standard error, is shown on the right-vertical axis. Because no pharmacogenetic interaction is simulated in Scenario 1, the statistical power in Figure 1 is shown with respect to the simulated SNP1 effect of $+2$ mmHg (i.e. $\beta_3=2$) per copy of the minor allele only. In Figures 2 and 3, the statistical power is shown with respect to the simulated SNP1 effects of both $+2$ and -2 mmHg per copy of the minor allele (i.e. $\beta_3=2$ and $\beta_3=-2$, respectively). Mean bias is displayed in all figures with respect to $\beta_3=2$. Figures 1–3 are based on 1000 simulation runs for each scenario. This seems to be a sufficient number of runs, since we obtained similar results with 10 000 simulation runs (data are not shown here).

The approaches to analysis are arranged across the x -axis in group order: *Naïve* [approaches (a)–(c)], *Substitution* [approaches (d)–(g)], and *Informative BP* [approaches (h)–(j)]. Results for an additional analysis are also included in each figure, for comparison purposes, on the far-left of the x -axis. This is

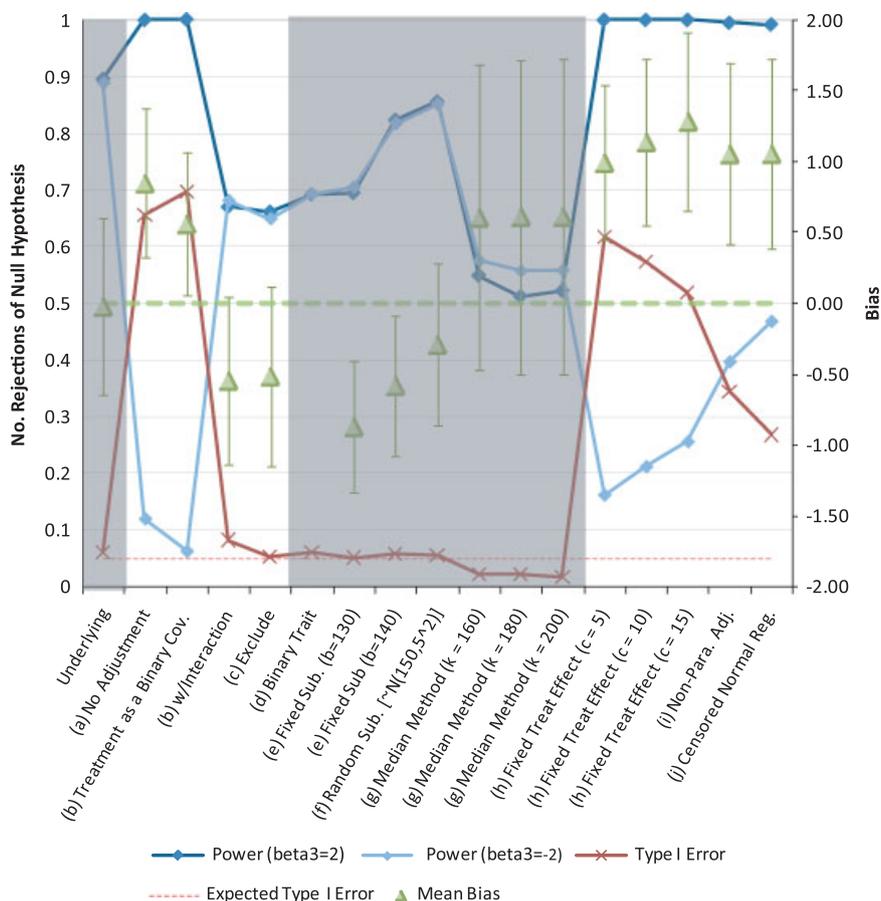


Figure 2. Results of Scenario 2—'Extreme Pharmacogenetic Interaction'. Approaches are arranged in categories (from left to right): *Naïve*, *Substitution*, *Informative BP*, highlighted by the shaded regions. The left-most approach is the analysis of underlying SBP, which is included for comparison purposes. Power is denoted in blue and light blue diamonds (with respect to the simulated SNP1 effects of +2 and -2, respectively); type I error (with respect to the simulated SNP1 effect of 0) is denoted in red crosses. Powers and levels of type I error are evaluated on the left-vertical axis, and lines joining the points are displayed for clarity. Mean bias (with SE) with respect to $\beta_3=2$ is displayed in green triangles, and is evaluated on the right-vertical axis (in mmHg). Because (d) fits a logistic regression and yields log-odds ratios, estimates of the effect of SNP1 are not comparable to the other approaches and are omitted from the plot.

an analysis of the underlying SBP, which, in practice, would not be observable for all participants due to the use of antihypertensive treatments. The analysis of underlying SBP demonstrates the maximum level of performance reasonably attainable given the simulation characteristics (i.e. the simulated sample size, the magnitude of the SNP effect, etc.). Because the analysis of underlying SBP is not affected by the use of treatment, the results are the same in all three scenarios, as would be expected. In Scenarios 2 and 3, a further additional analysis is also performed. This analysis is an extension of approach (b), modelling *treatment* as a binary covariate, but with a *SNP1-treatment* interaction term explicitly included too. Results for this additional analysis are therefore represented in Figures 2 and 3 adjacent to the results for (b).

Scenario 1 (Figure 1): The analysis of underlying SBP shows that, given the simulation characteristics, the maximum statistical power reasonably attainable in these simulations is approximately 90 per cent. The analysis of underlying SBP has the expected level of type I error (5 per cent), and, on average, is unbiased for the (main) effect of SNP1. Results most closely resembling those yielded by the analysis of underlying SBP are obtained by the Informative BP approaches [*Fixed Treatment Effect* (h), *Non-parametric adjustment* (i) and *Censored Normal Regression* (j)]. Each of the Informative BP approaches yields a high statistical power close to 90 per cent, the expected level of type I error, and only a small magnitude of bias (mean bias ≈ -0.25 to 0.2 mmHg). These analyses therefore seem reasonable approaches to correct for the use of antihypertensives. For *Fixed Treatment Effect* (h), the

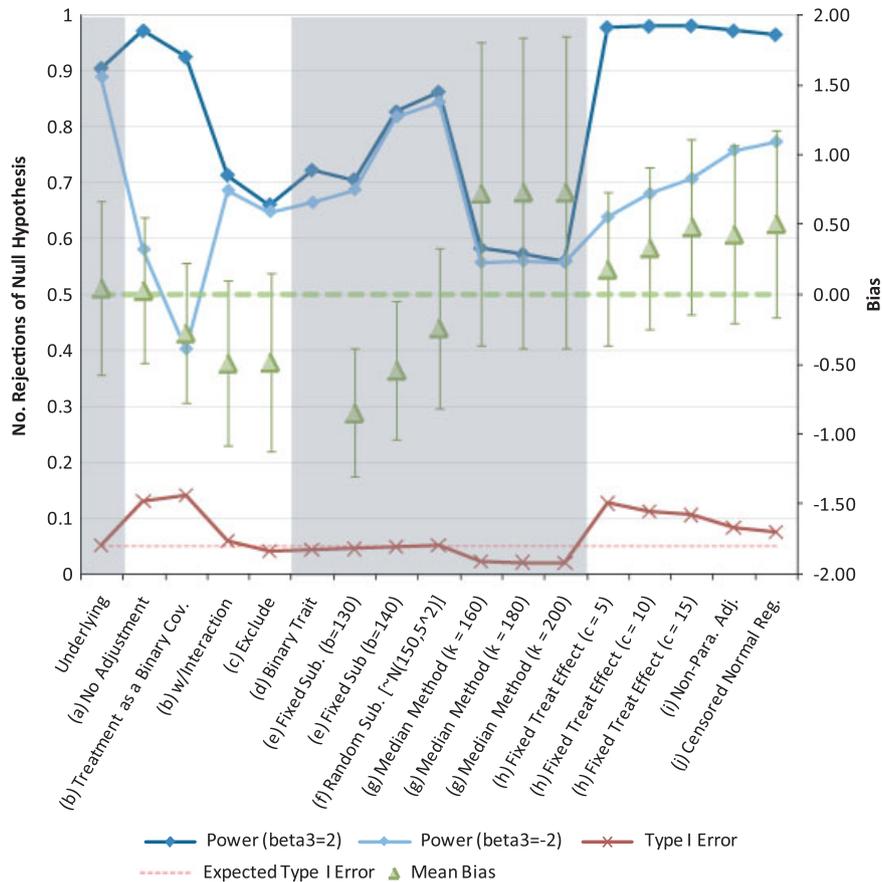


Figure 3. Results of Scenario 3—‘Pharmacogenetic Interaction with One Class of Antihypertensive’. Approaches are arranged in categories (from left to right): *Naïve*, *Substitution*, *Informative BP*, highlighted by the shaded regions. The left-most approach is the analysis of underlying SBP, which is included for comparison purposes. Power is denoted in blue and light blue diamonds (with respect to the simulated SNP1 effects of +2 and −2, respectively); type I error (with respect to the simulated SNP1 of 0) is denoted in red crosses. Powers and levels of type I error are evaluated on the left vertical axis, and lines joining the points are displayed for clarity. Mean bias (with SE) with respect to $\beta_3 = 2$ is displayed in green triangles, and is evaluated on the right-vertical axis (in mmHg). Because (d) fits a logistic regression and yields log-odds ratios, estimates of the effect of SNP1 are not comparable to the other approaches and are omitted from the plot.

results obtained are relatively stable with the different values of c tested (i.e. where c is the constant added to *modified BPs* to reverse the negative effect of treatment)—with decreasing levels of bias as c is closer to the simulated treatment effect. In agreement with previous findings, approach (h) thus seems relatively insensitive to the different choices of c used here, and performs well even when c differs considerably from the simulated treatment effect [20].

For the three *Naïve* approaches [*No Adjustment* (a), *Treatment as a Binary Covariate* (b), and *Exclude* (c)], estimates of the effect of SNP1 are shrunken towards the null (i.e. are closer to the null effect of zero), and there are consequent losses of statistical power. Hence, although the *Naïve* approaches do retain the correct level of type I error, it is clear that they are suboptimal strategies for analysis. With regard to the *Substitution* approaches [*Binary Trait* (d), *Fixed Substitution* (e), *Random Substitution* (f), and *Median Method* (g)], none of the results seem favourable in comparison to those of the *Informative BP* approaches. For instance, although the parameter coefficients for *Binary Trait* (d) cannot be compared with the other approaches because they are log-odds ratios (obtained by logistic regression), (d) has a very low statistical power (≈ 70 per cent). Likewise, although *Random Substitution* (f) yields reasonable power (≈ 87 per cent), a comparison with the similar *Fixed Substitution* (e) approach (power ≈ 70 – 83 per cent) shows that (e) and (f) are sensitive to the ‘substitution values’ used to replace *modified BPs*. In our simulations, we know that the value of 130 mmHg is below the threshold for initiating antihypertensive treatment. In practice, we will not always know the relevant threshold

and even when guidelines for treatment are widely available they are not always rigidly adhered to [38]. As the substitution value is increased from 130 to 140 mmHg [with (e)] and to the mean value of 150 mmHg [with (f)], the mean bias decreases from -0.9 to -0.3 mmHg. Hence, although (e) and (f) can potentially perform reasonably, they are highly influenced by the choice of substitution value. The *Median Method* (g) also appears sensitive to a substitution parameter (i.e. to the constant k). When k is 140, the SNP1 effect is underestimated (mean bias ≈ -1 mmHg), and when k is 150 or 160, SNP1 is overestimated (mean bias ≈ 0.2 – 0.7 mmHg). White *et al.* [19] state that (g) should actually be *insensitive* to k so long as the value chosen for k is sufficiently large. In subsequent scenarios, greater values for k are therefore chosen. Nevertheless, in this scenario, with any of the three values of k tested, (g) yields a low statistical power (≈ 47 – 60 per cent) and a lower level of type I error than expected (≈ 1.5 per cent), despite the fact that there were never more than 50 per cent of individuals on treatment in any simulation run.

Scenario 2 (Figure 2): In this scenario an extreme pharmacogenetic interaction is implemented, and the effects upon two of the Naïve approaches [*No Adjustment* (a) and *Treatment as a Binary Covariate* (b)] and each of the Informative BP approaches [*Fixed Treatment Effect* (h), *Non-parametric adjustment* (i) and *Censored Normal Regression* (j)] are striking. For instance, each of these approaches now markedly overestimates the marginal effect of SNP1 [mean bias ≈ 0.5 mmHg for (a) and (b); mean bias ≈ 0.5 to 1 mmHg for (h), (i), and (j)], and the power and type I error rates of these approaches are thus affected. When the simulated SNP1 effect is 2 mmHg ($\beta_3 = 2$), the power for each of these approaches is increased (≈ 100 per cent), but when $\beta_3 = -2$ the powers are reduced substantially [power ≈ 10 per cent for (a) and (b); power ≈ 18 – 48 per cent for (h), (i) and (j)]. The type I error rates for these approaches are also similarly affected. In contrast to Scenario 1—where each approach yields the correct level of type I error, the type I error rates in this scenario are highly elevated [type I error ≈ 0.7 for (a) and (b); type I error ≈ 0.3 – 0.6 for (h), (i), and (j)]. Given that these approaches do not account for what is effectively a *SNP1-treatment* interaction, this pattern of results is not unexpected. A reduced treatment effect is associated with the minor allele at SNP1. Ignoring the true effect of SNP1, individuals homozygous for (possessing two copies of) the minor allele who receive antihypertensive treatment will be less responsive to the treatment and will, on average, have greater *modified BP* than treated individuals who are homozygous for the major allele. Hence, the estimates of the main effect of SNP1 are biased upwards for these approaches in this scenario.

In contrast, *Exclude* (c) and the Substitution approaches [*Binary Trait* (d), *Fixed Substitution* (e), *Random Substitution* (f), and *Median Method* (g)] perform similarly in this scenario as in Scenario 1. These approaches avoid the effect of any differences in treatment efficacy between individuals because they replace *modified BPs* with alternative values derived independently from the observed data [or, in the case of (c), exclude any *modified BPs* from the analysis]. Hence, each of these approaches is unaffected by the pharmacogenetic interaction simulated with SNP1. These approaches are criticized in Scenario 1, however, and the same criticisms also apply here. For instance, (c) and (d) are low powered approaches, whereas (e) and (f) seem highly sensitive to the substitution values used. For (g), although the results are now stable with the different values of k implemented (because greater values of k are used in this scenario compared with those used in Scenario 1), a low power is yielded and there is a large magnitude of bias.

The additional analysis performed in this scenario, which models the *SNP-treatment* interaction term in addition to the *treatment* main effect, yields similar results here to *Treatment as a Binary Covariate* (b) in Scenario 1. Thus, although this approach is unaffected by the pharmacogenetic interaction, its estimates of the marginal effect of SNP1 are, on average, shrunken to the null (i.e. closer to the null effect) and it has a low statistical power (≈ 70 per cent). This approach avoids the problems of some of the other approaches in this scenario because it explicitly models the *SNP1-treatment* interaction term. This is actually the only approach that can account for potential SNP-treatment interactions in this way, because, in order to do this, the *treatment* main-effect must first be fitted. For reasons that are to be discussed later, adjusting for treatment use by modelling *treatment* as a binary covariate is a flawed approach to the analysis of BP. This approach, thus, cannot be considered as an optimal approach to analysis.

Scenario 3 (Figure 3): The pharmacogenetic interaction implemented in this scenario affects only participants on Treatment A, and thus has a more moderate effect than the extreme pharmacogenetic interaction in Scenario 2. Hence, for the Informative BP approaches [*Fixed Treatment Effect* (h), *Non-parametric Adjustment* (i), and *Censored Normal Regression* (j)] and two of the Naïve approaches [*No Adjustment* (a) and *Treatment as a Binary Covariate* (b)], there is generally less bias in this scenario

than in Scenario 2, and the statistical powers and type error rates are less badly affected. Type I error rates remain above 5 per cent (≈ 8 –13 per cent), however, and when $\beta_3 = -2$, the statistical powers remain substantially lower than those obtained in Scenario 1 (≈ 40 –80 per cent).

As in the previous scenario, the Substitution approaches [*Binary Trait* (d), *Fixed Substitution* (e), *Random Substitution* (f), and *Median Method* (g)] and *Exclude* (c) are again completely unaffected by the pharmacogenetic interaction, and perform similarly here as in Scenario 1. The additional analysis that models the SNP1-*treatment* interaction is also unaffected, and again yields similar results here to (b) in Scenario 1.

5. Discussion

Rapid progress is being made identifying genetic variants associated with BP in large-scale genome-wide association studies [13, 14]. However, as yet unidentified genetic determinants of BP are likely to have even more modest effect sizes than those already discovered. Approaches to maximizing the statistical power therefore remain important, and the need for an appropriate approach to analysis—which controls type I error—remains vital.

5.1. Summary and explanation of the results

Our simulations show that when the intervention is non-differential (i.e. in Scenario 1), the best approaches to analysis are clearly the Informative BP approaches [*Fixed Treatment Effect* (h), *Non-Parametric Adjustment* (i), and *Censored Normal Regression* (j)]. The Informative BP approaches yield similar results to the optimal analysis of underlying BP in this setting, and they thus appear to adequately control for the use of treatment. This finding supports previous work [20], which recommended these approaches for analyses of BP. The Informative BP approaches exploit all the observed data within analyses, and they maintain the natural variability between BP measurements between individuals. It is for this reason that they perform well in Scenario 1 but, conversely, this also explains why these methods are badly affected when there is an interaction with treatment (such as in Scenarios 2 and 3).

Like the Informative BP approaches, the extension to *Treatment as a Binary Covariate* (b) that models the SNP1-*treatment* interaction term also utilizes all the observed data within analyses. However, it does not suffer from the bias yielded by the Informative BP approaches in Scenarios 2 and 3 because it accounts for the differences in treatment efficacy between individuals by fitting the interaction term. As stated earlier, this is the only approach that can easily account for a SNP-treatment interaction because it is the only model that includes *treatment* as a covariate. It is well established, however, that modelling *treatment* as a binary covariate is a flawed approach to analyses of BP [20]. For instance, because the use of antihypertensives in this setting both predicts BP and is a consequence of having high BP, treatment should not be handled as a conventional covariate. Doing so explains away variation within the data, and attributes this variation to an apparent ‘treatment effect’. Including a treatment main effect term within an analysis model can thus mask true causal factors of BP—such as genetic variants—which are usually the main focus of a study.

The Substitution approaches [i.e. *Binary Trait* (d), *Fixed Substitution* (e), *Random Substitution* (f), and *Median Method* (g)] and *Exclude* (c) are unaffected by the pharmacogenetic interactions implemented in Scenarios 2 and 3, but yield sub-optimal results. They are also typically highly sensitive to the values of the ‘substitution parameters’. Given these findings, there is no obvious choice of approach that can be expected to perform well in all situations, but some practical recommendations will now be discussed.

5.2. Practical recommendations

It could be argued that because the Substitution approaches successfully control the type I error rates in all our simulations, their use should be preferred to the Informative BP approaches. However, as we have already noted, inappropriate choices of the substitution values could severely limit an investigation in terms of its ability to detect any undiscovered genetic variants, which are suspected to have very small effect sizes. Although guidance on the choice of the substitution parameters is available (e.g. [17, 19, 30]), different values for these parameters will be better suited to different circumstances. In practice, it may be difficult to choose suitable values for these parameters—and, indeed, it would be difficult to verify how suitable existing choices are. A further limitation of the Substitution approaches

is that they rely on an assumption that all individuals who use antihypertensives are hypertensive. It is known that antihypertensive medications are also prescribed for other conditions such as CHD, heart failure and migraine; hence, this is clearly a strong assumption. Although this assumption was true in our simulations, it has been shown that the Substitution approaches perform poorly when it is violated [20]. For these reasons, we therefore do not generally recommend use of the Substitution approaches for a primary analysis.

In addition to the scenarios reported in this paper, we have also tested another scenario in which a pharmacogenetic interaction is simulated, where the focus of the analyses is on estimating the effect of a SNP that is independent of the interaction. This scenario has shown that estimates of the marginal effects of any independent genetic or non-genetic factors are unaffected if another SNP is involved in a pharmacogenetic interaction. Thus, although performance of the Informative BP approaches is affected in the presence of a pharmacogenetic interaction with the genetic variant of interest, they would still appear to be the best approaches to use to estimate main effects in the absence of such an interaction. Note that estimates of the effects of genetic variants that are correlated (i.e. in *linkage disequilibrium*) with a SNP involved in a pharmacogenetic interaction will also be affected using these approaches. In an ideal world, it would be possible to identify *a priori* (from published pharmacogenetic studies of BP) those SNPs likely to be involved in differential treatment effects. Tests of the marginal effects of these particular SNPs could then be performed with an approach immune to the effects of a pharmacogenetic interaction (such as one of the Substitution approaches), whereas tests of all remaining, independent SNPs could use the Informative BP approaches. To date, although there is strong evidence of a genetic component to the variability of BP responses to antihypertensives, findings identifying loci for specific pharmacogenetic interactions with antihypertensives have not replicated [25, 26, 39]. Consistent published evidence for these effects is therefore currently lacking. Nevertheless, it seems reasonable to assume that only a small proportion of genetic variants across the human genome will alter the efficacy of antihypertensive treatments. Given this assumption, we therefore recommend that primary analyses of BP—which aim to detect SNPs that have an effect on underlying BP—should be performed using the Informative BP approaches. However, due to the lack of information about which regions of the genome have discernable effects on underlying BP and also alter the efficacy of antihypertensives, we would advise a critical interpretation of such results. In particular, pending further firm biological evidence about pharmacogenetic interactions, there may be exploratory analyses that could be undertaken with the data set under study to provide insight about potential interactions with the SNP of interest. Although not recommended as a primary analysis, one option to investigate the possible presence of a pharmacogenetic interaction for the SNP of interest would be to use the extended analysis of *Treatment as a Binary Covariate* (b) [i.e. which models the *SNP-treatment* interaction term]. Interactions are generally detected at a lower power than main-effects, however, and extensive follow-up work will be required to clarify whether such an approach would be reliable. An alternative would be to compare findings from an Informative BP approach and one of the Substitution [or *Exclude* (c)] approaches. The latter approaches are unaffected by pharmacogenetic interactions. If the results from the two analyses do not differ, it may be reasonable to assume that no strong pharmacogenetic interaction is present. However, further work is required to illustrate how large a discrepancy between the findings of these different approaches might be expected for real situations, as evidence about the characteristics of variants (e.g. minor allele frequency, main effect and interaction effect sizes, and directions) involved in pharmacogenetic interactions becomes available.

5.3. Implications

Our study shows that otherwise sensible approaches to the analysis of BP are affected when a genetic variant of interest influences treatment efficacy. Estimates of the marginal effects of genetic variants involved in pharmacogenetic interactions may therefore be biased—possibly leading to false-negative and false-positive findings. Pharmacogenetic interactions can thus impact on the statistical power of a study and on the level of type I error.

In principle, our results suggest that reported findings from existing genetic association studies could contain errors as a result of pharmacogenetic interactions. For instance, a genetic variant that influences treatment efficacy could yield spurious association with BP, or, conversely, a genetic variant that truly influences BP could be masked if it is also involved in a pharmacogenetic interaction. A secondary aim of this work could be to characterize such cases. Although analyses such as dichotomous hypertension [*Binary Trait* approach (d)] are low powered for a primary analysis, they could provide useful subsequent

checks to help identify whether novel genetic associations could be driven by a pharmacogenetic effect. For instance, all the genetic variants reported by Newton Cheh *et al.* [13] were associated with dichotomous hypertension in addition to continuous SBP and DBP, and are therefore unlikely to be fallacious. The issue of type I error due to a pharmacogenetic interaction is thus unlikely to be a problem in this particular study. However, the possibility of *type II error* remains. In addition to the strength of the interaction and the number of individuals involved, type II error will also depend on the direction of the interaction in relation to the direction of the main effect.

5.4. Applicability of our findings

We used simulation to demonstrate the potential influence of pharmacogenetic interactions in analyses of BP because, in practice, the true model generating mechanism is unknown. Furthermore, as yet, there is little known regarding the true nature and magnitude of pharmacogenetic interactions with antihypertensives. The actual influence of pharmacogenetic interactions in real analyses of BP is thus difficult to determine. For instance, if particular genetic variants interact with multiple classes of antihypertensive, there is a potential for serious distortions of the data (such as those shown in Scenario 2), but if pharmacogenetic interactions are specific to particular classes of antihypertensive, the implications could be less drastic (such as in Scenario 3).

Until now, we have focussed only on the analysis of BP in this paper, but our findings are also relevant to the analysis of other traits. For example, cholesterol-lowering drugs are widely used within western countries, and the investigation of low-density lipoprotein (LDL) and high-density lipoprotein (HDL) may thus also require one of the corrections for treatment described. Notably, because a single class of treatment—statin therapy—is predominantly used to lower cholesterol, any pharmacogenetic interaction would most likely apply to the majority of subjects on treatment. Hence, although the situation we simulated in Scenario 2 could be considered extreme for a study of BP, it may, in fact, be quite typical of a study of LDL/HDL.

In this paper we have focussed on the influence of a differential treatment effect, as this is the most likely form of a differential intervention in genetic studies. However, we have also considered (in unpublished work) the influence of a differential threshold for receiving treatment. Ultimately, both forms of a differential intervention lead to similar conclusions. For instance, estimation of the parameter that modifies either the treatment effect or the threshold for receiving treatment is often distorted, but estimation of all other parameters is generally unaffected. Hence, if the ‘modifying parameter’ itself is known but is not of interest, analyses may be performed without regard to our findings; however, when the modifying parameter needs to be estimated (and may or may not be unknown), difficulties may arise. Although we have suggested possible approaches to verifying results from genetic analyses of BP and to identifying potential pharmacogenetic interactions, further work is clearly required in these areas.

6. Conclusions

We suggest that the Informative BP approaches remain the most reasonable approaches to use for primary analyses of the main effects of SNPs in most settings. Nevertheless, caution is required in the interpretation of any associations obtained from these approaches. If there is strong *a priori* evidence of a particular pharmacogenetic interaction, it makes sense to consider the results of a different approach for the particular genetic variant involved. As further evidence of the nature and magnitude of pharmacogenetic interactions with BP emerges, a more detailed examination of the various approaches, their comparability, and possible methods for checking for these interactions will be warranted.

Acknowledgements

This work was supported by a British Heart Foundation Studentship (FS/06/040) to Nick Masca. Nuala Sheehan is supported by the Leverhulme Trust (Research Fellowship RF/9/RFG/2009/0062) and the Medical Research Council (Project Grant G0601625). Martin Tobin is supported by the Medical Research Council (Clinician Scientist Fellowship G0501942).

References

1. Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. *The Lancet* 2005; **365**:217–223.
2. Murray CJL, Lopez AD. Mortality by cause for eight regions of the world: global burden of disease study. *The Lancet* 1997; **349**:1269–1276.
3. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo Jr JL, Jones DW, Materson BJ, Oparil S, Wright Jr JT. Seventh report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension* 2003; **42**:1206.
4. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Prospective studies collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet* 2002; **360**:1903–1913.
5. Beilin LJ. Stress, coping, lifestyle and hypertension: a paradigm for research, prevention and non-pharmacological management of hypertension. *Clinical and Experimental Hypertension* 1997; **19**:739–752.
6. Pickering TG. The effects of environmental and lifestyle factors on blood pressure and the intermediary role of the sympathetic nervous system. *Journal of Human Hypertension* 1997; **11**:S9.
7. Havlik RJ, Garrison RJ, Feinleib M, Kannel WB, Castelli WP, McNamara PM. Blood pressure aggregation in families. *American Journal of Epidemiology* 1979; **110**:304.
8. Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham heart study. *Hypertension* 2000; **36**:477–483.
9. Lifton RP, Gharavi AG, Geller DS. Molecular mechanisms of human hypertension. *Cell* 2001; **104**:545–556.
10. Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene–environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology* 2003; **32**:51–57.
11. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *International Journal of Epidemiology* 2009; **38**:263–273.
12. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 2008; **9**:356–369.
13. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S, Papadakis K, Voight BF, Scott LJ, Zhang F, Farrall M, Tanaka T, Wallace C, Chambers JC, Khaw K-T, Nilsson P, van der Harst P, Polidoro S, Grobbee DE, Onland-Moret NC, Bots ML, Wain LV, Elliott KS, Teumer A, Luan Ja, Lucas G, Kuusisto J, Burton PR, Hadley D, McArdle WL, Brown M, Dominiczak A, Newhouse SJ, Samani NJ, Webster J, Zeggini E, Beckmann JS, Bergmann S, Lim N, Song K, Vollenweider P, Waeber G, Waterworth DM, Yuan X, Groop L, Orho-Melander M, Allione A, DiGregorio A, Guarrera S, Panico S, Ricceri F, Romanazzi V, Sacerdote C, Vineis P, Barroso I, Sandhu MS, Luben RN, Crawford GJ, Jousilahti P, Perola M, Boehnke M, Bonnycastle LL, Collins FS, Jackson AU, Mohlke KL, Stringham HM, Valle TT, Willer CJ, Bergman RN, Morken MA, Doring A, Gieger C, Illig T, Meitinger T, Org E, Pfeufer A, Wichmann HE, Kathiresan S, Marrugat J, O'Donnell CJ, Schwartz SM, Siscovick DS, Subirana I, Freimer NB, Hartikainen A-L, McCarthy MI, O'Reilly PF, Peltonen L, Pouta A, deJong PE, Snieder H, van Gilst WH, Clarke R, Goel A, Hamsten A, Peden JF, Seedorf U, Syvanen A-C, Tognoni G, Lakatta EG, Sanna S, Scheet P, Schlessinger D, Scuteri A, Dorr M, Ernst F, Felix SB, Homuth G, Lorber R, Reffelmann T, Rettig R, Volker U, Galan P, Gut IG, Hercberg S, Lathrop GM, Zelenika D, Deloukas P, Soranzo N, Williams FM, Zhai G, Salomaa V, Laakso M, Elosua R, Forouhi NG, Volzke H, Uiterwaal CS, van der Schouw YT, Numans ME, Matullo G, Navis G, Berglund G, Bingham SA, Kooner JS, Connell JM, Bandinelli S, Ferrucci L, Watkins H, Spector TD, Tuomilehto J, Altshuler D, Strachan DP, Laan M, Meneton P, Wareham NJ, Uda M, Jarvelin M-R, Mooser V, Melander O, Loos RJF, Elliott P, Abecasis GR, Caulfield M, Munroe PB. Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genetics* 2009; **41**:666–676.
14. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, Aulchenko Y, Lumley T, Kottgen A, Vasan RS, Rivadeneira F, Eiriksdottir G, Guo X, Arking DE, Mitchell GF, Mattace-Raso FUS, Smith AV, Taylor K, Scharpf RB, Hwang S-J, Sijbrands EJG, Bis J, Harris TB, Ganesh SK, O'Donnell CJ, Hofman A, Rotter JJ, Coresh J, Benjamin EJ, Uitterlinden AG, Heiss G, Fox CS, Witteman JCM, Boerwinkle E, Wang TJ, Gudnason V, Larson MG, Chakravarti A, Psaty BM, van Duijn CM. Genome-wide association study of blood pressure and hypertension. *Nature Genetics* 2009; **41**:677–687.
15. Wilcox J. Observer factors in the measurement of blood pressure. *Nursing Research* 1961; **10**:4–17.
16. Petrie JC, O'Brien ET, Littler WA, De Swiet M. Recommendations on blood pressure measurement. *British Medical Journal (Clinical Research edn)* 1986; **293**:611.
17. White IR, Chaturvedi N, McKeigue PM. Median analysis of blood pressure for a sample including treated hypertensives. *Statistics in Medicine* 1994; **13**:1635–1641.
18. Cook NR. An imputation method for non-ignorable missing data in studies of blood pressure. *Statistics in Medicine* 1997; **16**:2713–2728.
19. White IR, Koupilova I, Carpenter J. The use of regression models for medians when observed outcomes may be modified by interventions. *Statistics in Medicine* 2003; **22**:1083–1096.
20. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine* 2005; **24**:2911–2935.
21. McClelland RL, Kronmal RA, Haessler J, Blumenthal RS, Goff DCJ. Estimation of risk factor associations when the response is influenced by medication use: an imputation approach. *Statistics in Medicine* 2008; **27**:5039–5053.
22. Cui J, Harrap S. Genes and family environment explain correlations between blood pressure and body mass index. *Hypertension* 2002; **40**:7–12.

23. Cui J, Harrap S. Antihypertensive treatments obscure familial contributions to blood pressure variation. *Hypertension* 2003; **41**:207–210.
24. Raskin P. Treatment of hypertension in adults with diabetes. *Clinical Diabetes* 2003; **21**:120–121.
25. Turner ST, Schwartz GL, Chapman AB, Hall WD, Boerwinkle E. Antihypertensive pharmacogenetics: getting the right drug into the right patient. *Journal of Hypertension* 2001; **19**:1–11.
26. Arnett DK, Claas S, Lynch A. Has pharmacogenetics brought us closer to ‘personalized medicine’ for initial drug treatment of hypertension? *Current Opinion in Cardiology* 2009; **24**:333–339.
27. Matsubara M, Kikuya M, Ohkubo T, Metoki H, Omori F, Fujiwara T, Suzuki M, Michimata M, Hozawa A, Katsuya T. Aldosterone synthase gene (CYP11B2) C-334T polymorphism, ambulatory blood pressure and nocturnal decline in blood pressure in the general Japanese population: the Ohasama Study. *Journal of Hypertension* 2001; **19**:2179.
28. Charchar FJ, Tomaszewski M, Padmanabhan S, Lacka B, Upton MN, Inglis GC, Anderson NH, McConnachie A, Zukowska-Szczechowska E, Grzeszczak W. The Y chromosome effect on blood pressure in two European populations. *Hypertension* 2002; **39**:353.
29. Brand E, Wang JG, Herrmann SM, Staessen JA. An epidemiological study of blood pressure and metabolic phenotypes in relation to the Gbeta3 C825T polymorphism. *Journal of Hypertension* 2003; **21**:729–737.
30. Hunt S, Atwood L, Pankow J, Province M, Leppert M. Genome scans for blood pressure and hypertension—The National Heart, Lung, and Blood Institute Family Heart Study. *Hypertension* 2002; **40**:1–6.
31. Koenker R. quantreg: Quantile Regression, R package version 4.17, 2008. Available from: <http://www.r-project.org> [2008].
32. Buuren Sv, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**:681–694.
33. Cook NR. Imputation strategies for blood pressure data nonignorably missing due to medication use. *Clinical Trials* 2006; **3**:411–420.
34. Wang J, Kuusisto J, Vanttinen M, Kuulasmaa T, Lindström J, Tuomilehto J, Uusitupa M, Laakso M. Variants of transcription factor 7-like 2 (TCF7L2) gene predict conversion to type 2 diabetes in the Finnish Diabetes Prevention Study and are associated with impaired glucose regulation and impaired insulin secretion. *Diabetologia* 2007; **50**:1192–1200.
35. Wu J, Oberman A, Lewis C, Ellison R, Arnett D. A summary of the effects of anti-hypertensive medications on measured blood pressure. *American Journal of Hypertension* 2005; **18**:935–942.
36. Wang YR, Alexander GC, Stafford RS. Outpatient hypertension treatment, treatment intensification, and control in Western Europe and the United States. *Archives of Internal Medicine* 2007; **167**:141.
37. Turner ST, Schwartz GL, Chapman AB, Boerwinkle E. WNK1 Kinase polymorphism and blood pressure response to a thiazide diuretic. *Hypertension* 2005; **46**:758–765.
38. Wolf-Maier K, Cooper RS, Banegas JR, Giampaoli S, Hense H-W, Joffres M, Katarinen M, Poulter N, Primatesta P, Rodriguez-Artalejo F, Stegmayr B, Thamm M, Tuomilehto J, Vanuzzo D, Vescio F. Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States. *Journal Of the American Medical Association* 2003; **289**:2363–2369.
39. Arnett DK, Davis BR, Ford CE, Boerwinkle E, Leidencker-Foster C, Miller MB, Black H, Eckfeldt JH. Pharmacogenetic association of the angiotensin-converting enzyme insertion/deletion polymorphism on blood pressure and cardiovascular risk in relation to antihypertensive treatment: the genetics of hypertension-associated treatment (GenHAT) study. *Circulation* 2005; **111**:3374–3383.

Participant Identification in Genetic Association Studies: Improved Methods and Practical Implications

N. Masca, P.R. Burton and N.A. Sheehan*

Department of Health Sciences, University of Leicester, 2nd Floor Adrian Building, University Road, Leicester LE1 7RH, UK

*Corresponding author e-mail: Nuala.Sheehan@leicester.ac.uk

Summary

Background: In a recent paper (1), a method for detecting whether a given individual is a contributor to a particular genomic mixture was proposed. This prompted grave concern about the public dissemination of aggregate statistics from genome-wide association studies. It is of clear scientific importance that such data be shared widely, but the confidentiality of study participants must not be compromised. The issue of what summary genomic data can safely be posted on the web is only addressed satisfactorily when the theoretical underpinnings of the proposed method are clarified and its performance evaluated in terms of dependence on underlying assumptions.

Methods: The original method raised a number of concerns and several alternatives have since been proposed including a simple linear regression approach. In our proposed generalised estimating equation (GEE) approach, we maintain the simplicity of the linear regression model but obtain inferences that are more robust to approximation of the variance/covariance structure and can accommodate linkage disequilibrium.

Results: We affirm that, in principle, it is possible to determine that a ‘candidate’ individual has participated in a study, given a subset of aggregate statistics from that study. However, the methods depend critically on a number of key factors including: the ancestry of participants in the study; the absolute and relative numbers of cases and controls; and the number of SNPs.

Conclusions: Simple guidelines for publication that are based on a single criterion are therefore unlikely to suffice. In particular, *directed* summary statistics should not be posted openly on the web but could be protected by an internet-based access check (2).

Keywords: identification, linear regression, generalised estimating equations, linkage disequilibrium, case-control genetic association studies.

Word Count: 8,558

Box 1: Key Messages

1. We propose a more clearly justified method for participant identification in genetic studies.
2. When model conditions are satisfied, we affirm that reliable inferences are possible.
3. Our method is robust to correlation assumptions and does not require a specific model for LD.
4. Identification methods are sensitive to model assumptions and are hence study dependent

Introduction

Data collected in genetic epidemiological studies are, by nature, extremely sensitive and ensuring the protection of participant confidentiality is hence a matter of deep concern. Because of this, there are strict laws governing the sharing of individual-level genetic and non-genetic information (3). However, as advances in genomics research are informed and accelerated by the accessibility of results and summary information such as allele frequencies from genetic epidemiological studies, sharing of such aggregate data is often demanded by funding bodies (4, 5). Until recently, summary data from genome-wide association studies (GWAS) were freely available on the Web. However, in 2008, a statistical test was proposed that claimed to be able to detect an individual's presence in a DNA mixture from a genetic profile such as is typically obtained from a high-density single nucleotide polymorphism (SNP) genotyping platform (1). This was originally proposed with a forensics context in mind where the aim is to resolve whether a particular individual contributed DNA to a genomic mixture recovered from a crime scene, for example. However, the implications for genetic

association studies are obvious since the summary statistics of a study can be viewed as a mixture to which all participants contribute equally. Consequently, in view of the ethical and legal implications of even a potential violation of participant confidentiality, both the National Institute of Health and the Wellcome Trust felt compelled to alter their guidelines on the open web-based publication of summary information from genome-wide association studies (GWAS). The result is that aggregate data have been withdrawn from the internet and access has become restricted only to approved researchers (6). While the issues of consent and publication arising from this “blurring of traditional boundaries between individual and aggregate data” are still being debated, a number of interim recommendations have been made. These include a recent suggestion that no more than 500 regression results from any genetic association study should be published so that useful information on new findings and replication results can be provided without compromising the anonymity of an individual study participant (7). However, the real risks implied by methods such as described by Homer *et al* (1) are highly context specific and so it is important that the quantitative implications of the behaviour of the test are clearly understood before formulating any definitive recommendations. This is difficult, not least because the method outlined by Homer *et al* (1), and important variants such as that described by Visscher and Hill (8), are based on models which invoke assumptions that are often violated in real data.

This paper begins with a brief description of the test statistic originally proposed (1) and outlines some concerns about the theoretical underpinnings of that method. We then consider one particular alternative approach that has since been suggested (8) and examine some simple modifications that address the incorrect specification of the variance structure and the inevitable correlation due to linkage disequilibrium that exists in dense SNP data. We use simulations based on real data to assess how well our methods perform and how sensitive they are to the underlying assumptions. We conclude with a discussion of the implications of

our findings with regard to the inappropriate identification of subjects from case-control genetic association studies.

Methods

The Original Test Statistic

For consistency, we will use the following notation throughout. We have m SNPs in total with p_j denoting the true population minor allele frequency of the j th SNP. Our target is a DNA mixture or “test” sample from which we have obtained (minor) allele frequency estimates, \widehat{p}_j , for each SNP. We also have a “reference” sample with corresponding (usually estimated) allele frequencies, \widehat{p}_j^* . We have an individual, or *proband*, of particular interest (e.g. who could also be a suspect for a crime) for whom we have a full genetic profile. We want to know whether this individual is in our test sample. The called genotype for the proband at SNP j is denoted by g_j where $g_j \in \{0,1,2\}$ depending on whether there are 0, 1 or 2 copies of the minor allele at that SNP. Based on this single individual, the best estimate of the population minor allele frequency for the j th SNP is the observed frequency, $y_j = \frac{g_j}{2}$, $y_j \in \{0,0.5,1\}$. Under the assumption that the proband, test and reference samples share co-ancestry i.e. can be regarded as samples from the same overall population, we have that $E(y_j) = E(\widehat{p}_j) = E(\widehat{p}_j^*) = p_j$.

Note that the mixture, or test sample, frequencies, \widehat{p}_j , could be based on probe intensity values for a mixed DNA stain taken from a crime scene, as envisaged for the original forensic application (1). However, in the context of genetic association studies, they will be derived simply from the called genotypes of the individuals in the test sample. Similarly, the reference sample could be a conventional reference database, as would be used in a forensic setting, but will typically be just another sample in our applications, as will be discussed later.

The method proposed by Homer et al. (1) is defined on a SNP by SNP basis and is a simple comparison of two ‘statistical distances’: the distance between the individual of interest and the reference sample, and the distance between the individual and the test sample, or mixture. Thus for SNP j , the quantity of interest is given by

$$D_j = |y_j - \widehat{p}_j^*| - |y_j - \widehat{p}_j|. \quad (\text{Eq 1})$$

The basic idea is that over a large number of SNPs, the proband’s presence in the mixture, or test sample, will cause the values D_j to be positive, on average. Under the assumption of co-ancestry, the authors reason that an individual’s absence from the test sample would cause him to appear to be equidistant from both test and reference samples i.e. $E(D_j) = 0$ if absent. Assuming that D_1, \dots, D_m can be regarded as independent observations from a normal distribution with constant variance, combining information from all m SNPs leads to the conventional one-sample t-test statistic

$$T = \frac{\sum_{j=1}^m D_j}{\sqrt{\frac{\sum_{j=1}^m (D_j - \frac{\sum D_j}{m})^2}{m-1}}}. \quad (\text{Eq 2})$$

The proposed test of the null hypothesis that the proband is *not* in the test sample is the one-tailed test: $H_0: T = 0$ versus $H_1: T > 0$ (1). The authors claim that 10,000 to 50,000 SNPs are generally sufficient to be able to identify trace amounts of DNA (as little as 0.1 %) from an individual’s contribution to a complex mixture. Indeed, they argue that the number of SNPs required could be reduced significantly by careful selection of those SNPs that do not vary much between populations.

There was much confusion generated by the initial responses to this announcement and hence considerable uncertainty with regard to its implications for real data situations. This was

partly due to a rather heuristic exposition and lack of clarity in the original paper, both in the description and statistical underpinning of the proposed hypothesis test. Although the distributional assumptions assumed in (Eq 2) are never formally discussed, it would seem that the assumption of normality of the SNP distance measures, D_j , for a given individual is not unreasonable. However, it is unlikely that the D_j s will be independent for real SNP data. There are other problems with the proposed test. For one thing, the assumption underlying the one-sample test that $E(D_j) = 0$ under H_0 is questionable. If the proband is not in the test sample then, assuming co-ancestry, he should look like a typical member of the general population. Hence, under the null hypothesis and if the reference sample is reasonably representative of this population (as would be the case for frequencies derived from a reference database), his genetic profile will appear to be closer to the reference sample frequencies than to those of the test sample leading to an average *negative* mean distance, as can be deduced from (Eq 1). The original test thus has a composite null hypothesis under which the distribution of the test statistic (Eq 2) is not properly specified. This problem with the formulation of the null hypothesis is also discussed by Egeland et al (9) in their specific consideration of forensic mixtures. Moreover, the reference sample was only vaguely defined in the original paper and, indeed, the difference between the terms “sample” and “population” were altogether unclear (8, 10). Since population frequency estimates for genome-scan SNP data are unlikely to be reliable— or even available — it will be necessary to estimate these from another sample, in practice. The identification test is thus a two-sample problem where the proband can be in the test sample, the reference sample, or neither, and a two-tailed test is hence more appropriate. This point was also noted by Jacobs et al (11).

The two sample setting provides SNP allele frequencies \hat{p}_j and \hat{p}_j^* which are both estimates of the true population frequencies p_j . These estimates will differ in precision if the sample sizes of the two groups differ (12). In particular, the allele frequencies in the larger group will

generally be more representative of the population allele frequencies than those in the smaller group and, consequently, under the null hypothesis of being in *neither* test nor reference sample, the proband will appear to be “closer” to the larger group. It has been recommended that the reference sample should be bigger than the test sample (8, 11, 13). A one-tailed test would simply be conservative in this case whereas a two-tailed test would be biased towards inferring presence of the proband in the reference sample. A natural application of the method in a two-sample setting, and the focus of the rest of this paper, is a case-control genetic association study where the case and control groups can be regarded as deriving from one common population and can hence be tested directly against each other without the need for any additional reference sample (8, 10-12). However, since it is not reasonable to assume that case and control groups would always be equal in size, an approach to the identification problem that does not rely on this assumption is required.

It has been shown recently that the test given in (Eq 2) does not perform as originally claimed for “unbalanced” forensic mixtures where contributors are not equally represented and that reliable inference is impossible in these applications (9). However, the basic idea of Homer et al. (1) that summary statistics, such as allele frequencies, can identify the presence of an individual, or a close relative, in a genetic association study where DNA contributions are equal by design, appears to be surprisingly sound and identifies a real problem that warrants closer scrutiny (8, 10-14). In order to fully understand the implications of the identification issue, we need to be confident in the proposed statistical model and its underlying assumptions and be able to verify that it performs well when these assumptions are met. Only then can we assess how it performs when the assumptions are violated, particularly when such violations are likely to occur in practice.

An Alternative Approach

A more clearly motivated approach to the problem using linear regression was recently proposed which addresses some of the problems of the original test of Homer et al. while retaining the simplicity (8). We now outline the method and introduce some modifications to improve robustness to assumptions in realistic situations.

Linear regression

The question of interest is still whether the proband is in the mixture, or test sample. Instead of considering the two distance measures given by the absolute deviations between the proband's observed frequency at a particular SNP and the frequencies estimated from the test and reference samples, respectively, as given in (Eq 1), the idea is to regress the proband frequencies on those of the test sample where each is expressed as a deviation from the estimated population allele frequencies. Specifically, the model is

$$(y_j - \widehat{p}_j^{**}) = \beta(\widehat{p}_j - \widehat{p}_j^{**}) + \varepsilon_j \quad (\text{Eq 3})$$

where the ε_j are assumed to be independent and normally distributed with a constant variance. To allow for the fact that the reference sample may not be as representative of the overall population as might have been envisaged for the original forensic application of the one sample test in the previous section, the authors suggest that a pooled estimate, \widehat{p}_j^{**} , from both reference and test samples be used instead of the reference sample frequency, \widehat{p}_j^* , itself. A (one-tailed) chi-squared test is proposed to compare the hypotheses: H_0 : not in the test sample vs H_1 : in the test sample which again ignores the two-sample nature of the problem. Noting that the regression co-efficient, β , takes the value 0 under the null hypothesis that the proband is in neither the test nor reference sample, the value 1 if the proband is in the test sample and the value $-N_{test}/N_{ref}$ if the proband is in the reference sample where N_{test} and

N_{ref} are the sample sizes in the test and reference samples, respectively, we prefer the following two-tailed test

$$Z = \frac{\hat{\beta}}{\sqrt{\text{Var}(\hat{\beta}|out)}} \sim N(0,1) . \quad (\text{Eq 4})$$

Here $\text{Var}(\hat{\beta}|out) = \left(\frac{N_{test}}{m}\right) * \left(\frac{N_{test}+N_{ref}}{N_{ref}}\right)$ and by “out” we mean that an individual is in *neither* of the two groups. Regression methods have also been used to estimate the proportion of DNA contributed by an individual to a genomic mixture (15). In this case, testing for non-zero values of the appropriate regression co-efficient is a test for that individual’s presence in the mixture.

The proposed regression approach (8) performed well in a reported simulation study using independent SNPs and generally out-performed the approach of Homer et al. in terms of type I error rates and power to correctly detect the proband’s presence in the test sample. Importantly, it yielded approximately correct type I error rates regardless of the sample sizes of the two test groups. Co-ancestry between the mixture, the reference group, and the individual of interest is still a required assumption, as is independence of observations with constant variance. In practice, any of these assumptions could be violated. For example, although case-control GWAS are usually well matched in terms of ancestry, subtle differences between the two arms of a study could easily arise and would be difficult to detect. As for the test in (Eq 2), observations will often be correlated due to linkage disequilibrium (LD), and extracting a set of independent SNPS could lead to a substantial loss in power. Furthermore, the assumption about constant variances is unlikely to hold. This is easier to argue for the regression approach which is based on the scaled genotypes, $y_j = \frac{g_j}{2}$, rather than the distance measures, D_j , in (Eq 1). Assuming Hardy-Weinberg equilibrium (HWE), the genotypes, g_j , are binomial outcomes. Since the variance of a binomial random

variable depends functionally on the mean, the normal error distribution for the model in (3) is unlikely to be correct despite the fact that the deviations, $y_j - \widehat{p}_j^{**}$, are not themselves discrete. Note that Visscher and Hill (8) describe an alternative maximum likelihood approach which *does* model the error structure correctly under the stated assumptions but which also requires independent observations. We now propose two modifications of the linear regression approach to address the practical issues, firstly with regard to the error distribution and secondly, to allow for LD. We defer consideration of the co-ancestry assumption until later.

Modelling the variance

We consider two simple adaptations to the regression approach in order to address the variance misspecification issue. In the first case, we consider a logistic regression which solves the problem by modelling the variance function correctly. However, this still requires independent observations and so will not be expected to perform well in the presence of LD. In the second case, we maintain the simple linear formulation in (Eq 3) but consider it as a generalised estimating equation (GEE) with an independence structure. The idea here is to provide a more robust estimate of the standard error of the regression coefficient making inferences less sensitive to the variance misspecification. The added advantage is that GEEs are well suited to dealing with correlated data as we will discuss below.

Logistic Regression

Recall that under HWE the proband genotype for the j th SNP, $g_j \in \{0,1,2\}$ can be thought of as the number of “successes” (i.e. the number of copies of the minor allele) in two Bernoulli trials since an individual carries two alleles at each SNP. Thus

$$g_j \sim \text{Bin}(2, p_j), \quad j = 1, \dots, m$$

where p_j is the true, unknown, population minor allele frequency for SNP j . From (Eq 3), our explanatory variable of interest is $(\widehat{p}_j - \widehat{p}_j^{**})$ and so $p_j = E\left(\frac{g_j}{2} \mid (\widehat{p}_j - \widehat{p}_j^{**})\right)$ leading to the usual model for the log odds:

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta (\widehat{p}_j - \widehat{p}_j^{**}).$$

We adapt this to mimic the spirit of the linear regression model in (Eq 3) by adjusting for the (pooled) estimated population frequencies \widehat{p}_j^{**} with an offset on the log-odds scale:

$$\log\left(\frac{p_j}{1-p_j}\right) = \log\left(\frac{\widehat{p}_j^{**}}{1-\widehat{p}_j^{**}}\right) + \beta (\widehat{p}_j - \widehat{p}_j^{**}). \quad (\text{Eq 5})$$

Since the offset variable has a fixed co-efficient of 1, it can be used to adjust analyses without affecting the precision of other parameter estimates and essentially plays the role of a random effect here.

We fitted this model in R using the `glm()` package entering the outcome as two vectors: the first denoting the number of “successes” (i.e. g_j) and the second denoting the number of “failures” (i.e. $2 - g_j$). This is equivalent to fitting each genotype g_j as the sum of two binary allele variables, a_{1j} and a_{2j} (each with 0 or 1 copies of the minor allele), at a particular level of the covariate $(\widehat{p}_j - \widehat{p}_j^{**})$.

Generalised estimating equations

GEEs are usually used to provide consistent estimates of regression coefficients and their standard errors in correlated data and require some coherent ordering of the observations into *clusters* (16). Crucially, GEEs derive robust estimates of the standard errors of regression coefficients using the *sandwich* estimator of the variance via an iterative procedure which is consistent as long as the basic regression relationship is correct and there is no between-

cluster correlation in the responses (17). This holds even when the specified within-cluster correlation/covariance structure is incorrect. In this instance, we define the mean structure in the same way as the original linear regression model in (Eq 3):

$$(y_j - \widehat{p}_j^{**}) = \beta(\widehat{p}_j - \widehat{p}_j^{**}) + \varepsilon_j.$$

The model residuals, or estimated errors, $\widehat{\varepsilon}_j$, are then used to estimate the correlation parameters (given a pre-specified structure), before the model is refitted by applying an algorithm that incorporates these estimated correlation coefficients. This procedure is then iterated until the algorithm converges, i.e. when the estimates stabilise. The independence estimating equation is a special case where the correlation matrix is simply specified to be the identity matrix and a GEE model employed merely as a straightforward means by which to derive a robust estimate of the variance (18, 19). Hence, when every observation is in its own cluster, the independence GEE is similar to the linear regression method in (3), but uses the robust (or sandwich) estimator of the variance instead of the conventional variance. We fitted our GEE models in R using the `geepack()` function (20).

Accounting for linkage disequilibrium

Real GWAS datasets will have SNPs that are in LD and this correlation will be an issue if not properly accounted for. GEEs are especially useful when the nature of the correlation itself is not of primary interest, as they do not require the correlation matrix to be correctly specified. They thus seem particularly suitable for the identification problem where the correlation between SNPs can be considered as a nuisance factor.

We propose a first-order autoregressive (AR-1) correlation structure for a GEE model with the same mean structure as given in (Eq 3). In our context, this assumes that SNPs are correlated with other SNPs *within* a cluster and correlation strengthens with proximity.

Specifically, the correlation between two SNPs with a cluster ordering of j and k , respectively, is modelled as $\rho^{|j-k|}$. SNPs in one cluster are assumed to be uncorrelated with SNPs in a different cluster. An optimal choice for clustering SNP data would be to group SNPs by chromosome. However, at least 50 clusters are typically recommended (21), and there are computational limitations on the cluster size for simulations in which a new model is fitted for every test. For these reasons, we used clusters of size 20 here but acknowledge that the fitted correlation structure is only an approximation of the true LD structure between SNPs. Crucially, the robustness provided by the sandwich estimator of the variance ensures that although the resulting coefficient estimates may be inefficient, they should be consistent. The impact from any misspecification of the within-cluster correlation structure should therefore be limited. The influence of cluster size is discussed further below.

Results

Simulations

We begin with genotype data from the 1958 Birth Cohort (1958BC) (22) provided by the Wellcome Trust Case Control Consortium (WTCCC) (23). The 1958BC consists of 1,504 unrelated participants born in 1958 from twelve different regions of the UK, including Scotland and Wales but excluding Northern Ireland. Genotypes are typed on the Affymetrix 500K chip and called in Chiamo-Oxford format. For any individual's record, genotypes called with a probability of less than 0.9 are omitted. Furthermore, following advice in the exclusion files provided, 24 participants and 30,956 SNPs across all individuals were completely omitted.

We simulated case-control genetic association study data using the individual-level data from the 1958BC. In each simulation run, we randomly sampled (without replacement) one

hundred participants into each arm of a hypothetical study (i.e. the case and control groups) and another one hundred individuals into an additional group to test under the null hypothesis of “neither group”. For the first three analyses, we only used individuals from the southern UK regions (i.e. London, southeast, southwest or south England) and would thus expect the co-ancestry assumption to hold. There are 461 participants in total from which to sample with this restriction. In the fourth analysis, we investigate sensitivity to the co-ancestry assumption by introducing regional differences between the two arms of the case-control study. The fifth analysis is also concerned with co-ancestry but here we consider the more extreme case where the cases are drawn from a completely different genetic association study than the controls. For this scenario, we also used genotype data from the UK National Blood Service (NBS) (23) and the Coronary Artery Disease (CAD) (24) studies as provided by the WTCCC. Since real case-control studies are usually well-matched ancestrally, we argue that discrepancies between two different UK studies would be a reasonable reflection of any differences likely to be encountered in practice. The NBS comprises 1500 unrelated participants from the UK, and the CAD comprises 1988 unrelated, coronary artery disease patients also from the UK. As with the 1958BC data, all genotypes are typed on the Affymetrix 500K chip and called in Chiamo-Oxford format, and any genotypes called with a probability of less than 0.9 are omitted. We again followed the advice provided in the exclusion files and omitted 42 individuals from the NBS study, and 62 individuals from the CAD study. The same SNPs that were omitted from the 1958BC data were also omitted from these datasets.

Since strongly correlated data would mask any other effects due to ancestry or variance misspecification, we attempted to at least reduce the effects of LD in all our analyses by selecting a subset of SNPs evenly spaced across the genome. The wider the SNP spacing, the weaker the LD and so we constructed three datasets for the 1958BC consisting of genotypes

for every 20th, every 33rd and every 100th SNP, in genomic order, across chromosomes 1 to 22. For compatibility of findings, each analysis uses the same number of SNPs. Thus, since we have 4577 SNPs in total for the sparsest SNP spacing of 100, we only used the first 4577 SNPs (sorted by chromosome and position) for the other two spacings. Note that by ignoring so much data, we can expect to lose statistical power so all our results will be conservative.

For our purposes, the case group is generally the mixture, or test sample, and the control group is the reference sample, but these can, of course, be reversed. In every simulation run, we considered each individual from the case group, the control group, and from neither group in turn as a proband to be tested for presence in the overall genetic association study. We discuss results for a 5% significance level throughout since we do not have sufficient power to detect anything at genome-wide significance level with this number of SNPs. If a particular individual of interest has a genotype missing for a given SNP, that SNP is omitted from the corresponding test – although it may be included in other tests. We performed one thousand simulation runs for each analysis and derived mean estimates of the regression coefficient, β , and its variance, as well as Monte Carlo estimates of the power to detect the proband's presence in the overall study (i.e. in either the case or control group) and the type I error rate.

Analysis One – Linear regression Approach

The focus of this first analysis was to investigate the performance of the original linear regression approach (section 3.1) on the three datasets with SNP spacings of 20, 33 and 100, respectively. Results in Table 1 indicate that, on average, the regression coefficient, β , is estimated virtually without bias. (Note that since our two samples are the same size, $\beta = -1$ if the proband is in the reference sample.) Moreover, the power to infer the proband's presence in the study (i.e. the % rejections of H_0 for cases or controls) is consistently over

98%. However, the type I error rates, given by the percentage of rejections of H_0 for individuals who were in *neither* group, are all higher than the expected 5% level. These decrease with increasing SNP spacing but remain unacceptably high (about 8%) even when the spacing is 100. Since LD is unlikely to be a big problem when we take every 100th SNP, the elevated type I errors must be due to some other factor. An obvious candidate cause would be misspecification of the variance function. This is supported by the fact that when we simulated data with truly independent SNPs (not shown), the type I errors were still consistently slightly higher than expected at around 6%. This trend was also evident in the simulation results reported by Visscher and Hill (8) but was not explored there, presumably because the type I errors were much more stable than those for the test in (2) with which the method was being directly compared. We also noticed similar elevations in type I errors at the .0001 significance level for our own independent SNP scenario.

Table 1 Here

Analysis Two – Correcting the variance

In order to eliminate the impact of LD as effectively as possible, we restricted attention to the dataset with a SNP spacing of 100 although we would generally not recommend the discarding of so much data. Since we are still using 1958BC individuals from the southern UK regions, we can also assume that the co-ancestry condition holds. If the elevated type I error rates obtained for the same dataset in the previous analysis were due principally to misspecification of the variance function, we would expect improved performance from both the GEE independence model (because it is robust to misspecification of the covariance structure) and from the logistic regression model (because the functional relationship between variance and mean is correct) by comparison with the linear regression model.

Table 2 compares linear regression with logistic regression and the GEE independence model, where the latter was fitted with a cluster size of 1 i.e. every observation is in its own cluster. As can be seen, both alternative approaches perform very similarly in terms of type I error rate and power. Furthermore, as hypothesised, both exhibit a type I error rate that is much closer to nominal than that of the linear regression model. Moreover, the power remains high ($\approx 98\%$) and is only marginally lower than that of the linear regression approach ($\sim 99\%$). Although estimates of the regression coefficient for the logistic regression approach cannot be compared with the other approaches (i.e. because they are log-odds ratios), the GEE independence model estimates are identical to those of the linear regression model since the only difference is the robust estimate of the variance which is increased by about 20% on average (from ~ 0.49 to ~ 0.62). For truly independent SNPs, type I error rates for the linear regression approach were consistently just over 6%, as noted earlier and in line with those reported in (8), whereas GEE independence and logistic regression models both yielded type I errors close to the nominal 5% level (simulation results not shown).

Table 2 here

Analysis Three – Accounting for linkage disequilibrium

In order to introduce more LD, we now compare the different approaches using the 1958BC dataset with a SNP spacing of 20. Despite the initial thinning out of the data, we have to allow for the fact that the observations are correlated. We used a GEE AR-1 model with a cluster size of 20. This allows for LD to prevail over a range of 400 SNPs in the full data set but we do not claim that this models LD – or, in particular, the correlation structure *within* the 20 spaced SNPs – in any biologically realistic way: any correlation between SNPs in *different* clusters remains unaccounted for in these analyses. However, it is undoubtedly much better than ignoring the correlation altogether.

Table 3 shows that the three approaches that assume independence all have increased type I error rates in these data. Although the logistic regression and GEE independence models improve on the linear regression method, they still yield type I errors of around 7%. In contrast, the GEE AR-1 provides acceptable levels. We note that there is a trade-off, however, in that power to detect a proband's presence in the study is a little lower, although still reasonable. The variance of the estimated regression co-efficient for the GEE AR-1 approach (~0.73) is around 20% greater than that for the independence model with cluster size of 1 (~0.61) and around 40% greater than that for the linear regression model (~0.50), reflecting the added correlation. It is important to note that a GEE independence model with cluster size of 20 (i.e. fitting an identity matrix for the within-cluster correlation structure but specifying that the correlation structure applies to clusters of size 20 rather than 1) gave almost identical results to the GEE AR-1 model above implying that a fairly realistic model for the clustering is more important than the proposed correlation structure itself. This is not unexpected given the inherent robustness of the sandwich estimator of the variance to misspecification of the correlation structure *within* blocks (17), but not *between* blocks. Larger SNP spacings gave similar results, as might be predicted, with the GEE AR-1 model (or GEE models with clusters of 20) maintaining approximately correct levels of type I error.

Table 3 here

Analysis Four – Effect of differences in ancestry

Because real case-control GWAS are usually well matched in ancestry, we now focus on the effects of possible, subtle differences in ancestry. Using the 1958BC we simulated hypothetical case and control groups of 100 individuals from different regions of the UK. In particular, our first analysis sampled cases from southern UK regions and controls from central UK regions (i.e. east, north midlands, midlands, and Wales); a second analysis

sampled cases from southern regions and controls from northern regions (i.e. northwest England, north England, east & west ridings of Yorkshire, and Scotland); and a third analysis sampled cases from central regions and controls from northern regions. In addition, we also sampled a group of 100 individuals from each region to test under the null hypothesis. We used the dataset with a SNP spacing of 20 in these analyses and, hence, because of the known LD in these data, we only show results for the GEE AR-1 approach using a cluster size of 20. Note, as before, that we obtained almost identical results for the GEE independence model with the same cluster size.

Table 4 demonstrates that any differences in ancestry between participants from different regions of the UK in the 1958BC do not affect the power to detect the proband's presence in the case-control study. This is supported by the fact that, as with all previous analyses, the regression coefficient is estimated, on average, without bias in this scenario and the GEE AR-1 approach yields acceptable levels of type I error and a high power which are comparable to those obtained in Analysis 3 (Table 3 above). These findings are consistent with those reported by Clayton (10).

Table 4 here

Analysis Five – Comparing Different Cohorts

To further explore the effects of possible differences in ancestry in case-control GWAS, hypothetical case and control groups are simulated by sampling individuals from three real, UK studies. Different studies have different target populations, different recruitment procedures, and could be subject to different biases. It is of interest to see whether any such between-study differences can cause problems for identification testing, even when the studies appear to be well matched in ancestry. As above, we simulated one hundred hypothetical case and control groups by sampling 100 individuals without replacement from

the 1958BC, NBS, and CAD studies, and an additional group of 100 individuals from each study to test under the null hypothesis. We again used a SNP spacing of 20 for this analysis and again provide results for the GEE AR-1 approach only.

Table 5 shows a similar pattern of results to those obtained for Analysis 4. The regression coefficient is, on average, estimated without bias, and the type I error rates – where individuals not present in a case/control group are incorrectly inferred as being in that group – are all approximately correct. This suggests that any subtle ancestral differences between individuals in different genetic association studies are not sufficient to make identification intractable. The GEE AR-1 approach, thus, seems to perform well in case-control GWAS data, where none of its model assumptions are seriously compromised. We note that these findings are in line with the reported associational analyses from the WTCCC (23) which also used different studies for case and control groups and found that both these and the more modest regional differences had negligible effects on their results.

Table 5 here

More substantial differences, however, are likely to cause problems. Visscher and Hill (8) noted that violation of the co-ancestry assumption becomes an issue when the divergence between test and reference samples, as measured by Wright's F_{ST} (25) gets close to a value of $1/2N_{test}$. From this, we should expect impaired performances of our models for scenarios such as we simulated above, with 100 cases and 100 controls, when $F_{ST} > .005$. Note that an F_{ST} value of .005 is considered extreme for values typically observed between different European populations (10, 26). There are various formulae for calculating F_{ST} (27, 28) and published F_{ST} values for real populations are not always consistent (26, 28-31). We used a version given by Cavalli-Sforza (28) and our results may hence differ slightly from those in

(8) if a different formulation was used. Specifically, divergence at the j th SNP between two populations, 1 and 2, with minor allele frequencies p_{1j} and p_{2j} respectively, is given by:

$$F_{ST} = \frac{Var(p_j)}{\bar{p}_j (1 - \bar{p}_j)}$$

where $\bar{p}_j = \frac{1}{2}(p_{1j} + p_{2j})$ and $Var(p_j) = (p_{1j} - \bar{p}_j)^2 + (p_{2j} - \bar{p}_j)^2$. An overall measure of F_{ST} is obtained by taking the mean F_{ST} values across all SNPs. Note that F_{ST} , as given above, measures divergence between the two underlying populations as it involves the true population allele frequencies. In practice, of course, it has to be estimated from the sample frequencies and we would therefore expect it to be sensitive to both the actual and relative sizes of the two samples.

We simulated 5,000 independent SNPs per individual for two populations with varying degrees of divergence and calculated detection power and type I error, as before, for the linear regression, GEE independence (with cluster size of 1) and the GEE AR1 (with clusters of 20) models. Table 6 shows some results for the situation when the case and control groups are equally sized with 100 individuals. Since we know the “true” allele frequencies in this case, we have calculated F_{ST} using both the true and the observed allele frequencies in order to assess performance in situations where more representative samples may be available.

Table 6 here

Considering the more realistic setting where only the sample frequencies are available, we can see that while type I error levels are acceptable (i.e. very slightly elevated for the linear regression model and about the desired 5% for the two GEE models) at the suggested threshold value of .005, they start to increase very quickly with increasing F_{ST} and they are already quite bad at about .006 with type I error rates between 12% and 14%. When the true

frequencies are used, performance is degraded long before this threshold is attained and we observed type I error rates ranging from 77% to 92% when the F_{ST} value was .0049. This implies that relatively small changes in ancestry could have a considerable effect when “good” estimates of population allele frequencies are available. The GEE models were more adversely affected than the simple regression approach for high levels of F_{ST} but were better when the co-ancestry assumption was not seriously violated. Since the data are truly independent here, the two GEE models behaved similarly, as would have been expected.

The threshold of $1/2N_{test}$ (based on observed frequencies), however, was not a good indicator when the test and reference samples were of different sizes. For a case group of 100 and control group of 500, for instance, we observed type 1 errors at around 25% for an F_{ST} value of .004. Not only is it difficult to “quantify the limits of identification in practical situations” (8) but it also seems difficult to provide a simple rule, such as a threshold based on F_{ST} , to indicate when the reference sample is sufficiently different from the test sample as to render an identification test futile.

The practical implications are that under ideal conditions, “good” population cohorts such as the 1958BC and the NBS, which provide representative estimates of population allele frequencies, enable reliable inferences to be drawn about a proband’s presence in the study using a full genetic profile. This is entirely in agreement with Clayton’s clarification of the role of the reference frequencies (10). However, the tests are extremely sensitive to the assumption of co-ancestry and estimated allele frequencies, in particular, will often not be good enough to raise serious concerns about identification. For extremely divergent test and reference samples, such as would be provided by the HapMap Yoruba and CEU populations with an F_{ST} of around 0.15-0.2 (26), our simulations indicate that inference is likely to be problematic however accurate the allele frequencies.

Discussion

Our findings are entirely consistent with those reported elsewhere (1, 8, 10-14). Despite some problems with the theoretical justification, the idea behind the test proposed by Homer et al. (1) raises important issues that are undoubtedly pertinent to genetic association studies. Under certain conditions it *is* possible, given no more than the summary distribution of dense genotypes across a study, to infer whether a given individual (whose full genetic profile is known) did, or did not, participate in that study. We have shown here that this can be done with high power under ideal conditions with as few as 5,000 SNPs and summary statistics based on 100 individuals and indeed we concur with the overall finding in (1) that DNA contributions of as little as 0.1 % (corresponding to test sample sizes of 1000 in our applications) can be detected with about 50,000 SNPs (results not shown). We have argued that the mathematical model proposed by Visscher and Hill (8) is more coherent than that in the original paper and out-performs the original method. Importantly, it does not require that the test and reference samples be of equal size when the underlying assumptions are satisfied. However, the regression approach yields consistently elevated type I errors due to the inherent Binomial nature of genotype data and hence misspecification of the variance function and has further problems when the observations are correlated because of LD. The latter is also a problem for the original method and the various proposed alternatives but has only been briefly alluded to in the literature. Empirical distributions of unmodified test statistics in the presence of LD have been considered (1, 11) but it is clear that LD either has to be modelled appropriately (10) or otherwise accounted for, such as we suggest here.

If we are to properly understand the implications of the Homer et al methods, and to determine what may or may not be ‘safe’ in terms of limited data release, it is critical that we fit models that *are* inferentially ‘well behaved’. Here, we have described models that can deal both with the variance misspecification and with the correlation. For (reasonably)

independent SNPs, the variance misspecification is satisfactorily addressed either by modelling it correctly using logistic regression or, more simply and with only a slight loss in power, by using the sandwich estimator of the variance - as in a GEE independence model - which renders regression parameters consistent even if the covariance is misspecified. The size of the clusters in the GEE approach was not important in this case and the model worked equally well with each observation in its own cluster. Hence, it would seem that getting the variance *right* is not as important as allowing for the fact that it *might be wrong*. More densely spaced SNPs caused type I error problems for both models but a GEE AR-1 model with a cluster size of 20 performed well when the SNP spacing was 20 with only a slight reduction in power to infer the proband's presence in the case-control study. Importantly, we obtained very similar results for the independence GEE model with the same cluster size. This implies that identifying the *regions* of LD is more important than modelling the *nature* of LD within these regions. We thus agree with Clayton (10) who also notes that LD cannot be ignored, but we would argue that it can actually be regarded as more of a nuisance factor and, in particular, we can avoid the added computational problem of using an external data set to model the correlation structure and then inverting a large sparse correlation matrix using least angle regression techniques (10). Since the data can never be assumed to be truly independent, we recommend that it is safer to use a GEE model and allow for LD by clustering the data and imposing a convenient correlation structure. We would also recommend that the data be thinned both to balance the trade-off between cluster size and losing data and to allow for the fact that LD can be quite far-ranging. Larger clusters were particularly problematic for our simulations due to the numbers of individuals we tested in each run. This would not be an issue in practice, of course, where there would only be one proband to test (rather than numerous probands as part of a large simulation exercise).

Recall that our reported results were all based on the first 4,577 of the 25,000 available SNPs at this spacing and are thus conservative. However, it is important to note that large numbers of SNPs are not required for reliable inference from realistic sample sizes under ideal conditions. Indeed, Visscher and Hill (8) suggest that the SNP number to test sample size ratio is about 6 for a nominal type I error rate of 0.05 and 80% power, and is about 50 for a type I error rate of .0001 95% power. Denser SNPs required larger cluster sizes for our GEE approach (results not shown) confirming that LD can range over several hundred SNPs in an Affymetrix 500K scan. For instance, we considered un-thinned SNP data by taking all 15,000 SNPs from chromosome 14 and obtained type I error rates of about 45% for the linear regression model compared with 23% for the GEE independence model with no clustering. However, with clusters of 400 we still had type I errors of between 8% and 9% for the GEE independence and AR1 models. Although the LD range is the same here as in the earlier case for clusters of size 20 on every 20th SNP, the degradation in performance is due to the fact that the increased LD from the denser spacing is creating greater dependence *between* the clusters. In this case, we rectified the problem by taking every 10th SNP and using a cluster size of 100. Note that our reported type I error rates are all slightly higher than the desired 5% level so it is clear that the structure of our clusters is still not quite right. Ideally, we should cluster SNPs by chromosome but this falls short of the recommended number (typically 50) of clusters to fit a GEE model and is computationally more intensive. One possibility is the use of an external data set, such as the HapMap, to inform our clustering by identifying regions of strong LD which should ideally be contained within a single cluster. More importantly, spacing between clusters is necessary as our methods are sensitive to violations of the assumption that there is no between-cluster correlation.

Now that we have models that are behaving appropriately in realistic situations, we can begin to think about quantifying their behaviour when other basic assumptions are violated. The

assumption of co-ancestry is crucial to the original method [1] and to all proposed variants. We agree with Visscher and Hill (8) that the only relevant violation occurs when the ancestries of the test and reference samples differ. Indeed, from simulation studies of independent SNP data where hypothetical cases and controls were sampled by ‘disease’ status defined by arbitrarily chosen numbers of ‘causal’ SNPs, we found that, as long as the assumption of common ancestry was valid, our robust regression-type approaches were insensitive to other differences between the two groups, regardless of the number of such causal SNPs. We have verified that when ancestry is ‘good’, such as one would expect when case and control groups are taken from different regions within a representative UK cohort or from different such cohorts across the UK, strong inferences about a proband’s participation in the overall study can be drawn. Thus, it would be possible to infer whether an individual suspected of a crime was in the study and in which group (case or control) in that study (thus providing information on disease status), using a genome-scan genetic profile of that individual and summary allele frequencies from the overall study or from both case and control groups, respectively. However, one has to be very confident about ancestry in order to make such inferences as even small differences between the test and reference samples can lead to greatly inflated type I errors and hence erroneous conclusions (see Table 6).

There are important implications for what data can be made freely available, and this is highly study-dependent. The power to detect a proband’s presence in a study increases with decreasing test sample size and increasing reference sample size [12] and the number of SNPs required to be informative also depends on these sizes. For example, when the case and control groups both were of size 100, we were able to detect that a proband was a case with high power from just 5,000 SNPs. Increasing the control group to 500 showed that just 2000 SNPs sufficed (with power of ~93%). The smaller the test sample, the easier it is to detect a single individual’s contribution but we are unlikely to get case-control studies with

fewer than 100 in either group. The important message is that simple rules such as “no more than X results” are almost certainly not sufficient to guarantee participant confidentiality. Given the two-sample nature of the testing problem, however, it is safe to say that any ‘directed’ results — such as signed p-values indicating which alleles are associated with the outcome — are potentially informative and should probably not be published at all. On the other hand, *unsigned* p-values could be released without risk (10) and as they can be of great value in a number of settings, we believe it would be helpful to exclude them from any embargo on aggregate statistics following the reaction to Homer *et al* [1].

Finally, despite the theoretically strong underpinning of the basic conclusions of Homer *et al*, we would argue that the strong reliance of the methods on the underlying assumptions particularly that of co-ancestry — renders the true level of forensic or ethical risk imposed on study participants rather small in many practical situations. Thus, we would support the suggestion (2) that directed study-wide summary statistics from genetic association studies could be protected in many cases by an internet-based access mechanism that simply checks that a potential user is a *bona fide* biomedical researcher in good standing, such as is achievable using contemporary technology (<http://www.gen2phen.org>). In particular, it would seem that the huge sample sizes that are currently being collected for genome-wide association analyses will render identification — from study-wide summary statistics alone — forensically unreliable (13). This is likely to be particularly true for the common situation where sample size is increased by pooling different studies and where the assumption of common ancestry will be less likely to hold. However, the potential risk has to be carefully considered on a case by case basis. This paper will hopefully become part of the information process by which new policies can be constructed that aim to better balance the scientific value of sharing data with the assurances of confidentiality given to study participants before such methods were available and aid in the design of consent forms for future studies.

FUNDING

This work was supported by the British Heart Foundation (Studentship FS/06/040), the Leverhulme Trust (Research Fellowship RF/9/RFG/2009/0062) and the UK Medical Research Council (Project Grant G0601625). It made use of data and samples generated by the 1958 Birth Cohort (<http://www2.le.ac.uk/projects/birthcohort>) under grant G0000934 from the Medical Research Council, and grant 068545/Z/02 from the Wellcome Trust. Genotyping was undertaken as part of the Wellcome Trust Case-Control Consortium (WTCCC) under Wellcome Trust award 076113 and a full list of the investigators who contributed to the generation of the data is available at www.wtccc.org.uk. The methodological program at the University of Leicester focusing on genetic statistics and large scale data harmonization and sharing is also supported under: P³G (the Public Population Project in Genomics) funded by Genome Canada and Genome Quebec; PHOEBE (#FP6-2006 – 518418) and BioSHaRE-EU (#FP7-2010-261433) grants from the European Framework Program; Wellcome Trust Supplementary Grant (086160/Z/08/A); and by the Leicester Biomedical Research Unit in Cardiovascular Science (NIHR). None of the funders had any role in the analyses and interpretation of the data or in the preparation, review or approval of the manuscript.

REFERENCES

1. Homer N, Szelling S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet.* 2008;4(8):e1000167.
2. P3G_Consortium, Church G, Heeney C, Hawkins N, de Vries J, Boddington P, et al. Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection. *PLoS Genet.* 2009;5(10):e1000665.
3. Lowrance WW, Collins FS. Identifiability in genomic research. *Science.* 2007;317(5838):600.
4. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics.* 2008;9(5):356-69.
5. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics [mdash] re-shaping scientific practice. *Nat Rev Genet.* 2009;10(5):331-5.

6. Couzin J. Genetic Privacy: Whole-Genome Data Not Anonymous, Challenging Assumptions. *Science*. 2008;321(5894):1278-.
7. Lumley T, Rice K. Potential for Revealing Individual-Level Information in Genome-wide Association Studies. *JAMA*. 2010;303(7):659.
8. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet*. 2009;5(9):e1000628.
9. Egeland T, Fonnelop AE, Berg PR, Kent M, Lien S. Complex mixtures: a critical examination of a paper by Homer et al. *Forensic Science International: Genetics*. 2011;In press
doi:10.1016/j.fsigen.2011.02.003.
10. Clayton D. On inferring presence of an individual in a mixture: a Bayesian approach. *Biostatistics*. 2010;11(4):661-73.
11. Jacobs KB, Yeager M, Wacholder S, Craig D, Kraft P, Hunter DJ, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*. 2009;41(11):1253-7.
12. Braun R, Rowe W, Schaefer C, Zhang J, Buetow K. Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS Genetics*. 2009;5(10).
13. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*. 2009;41(9):965-7.
14. Sampson J, Zhao H. Identifying Individuals in a Complex Mixture of DNA with Unknown Ancestry. *Statistical Applications in Genetics and Molecular Biology*. 2009;8(1):37.
15. Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*. 2001;46(6):1372-8.
16. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.
17. Zeger SL, Liang KY. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*. 1992;11(14-15):1825-39.
18. White H. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*. 1982;50(1):1-25.
19. Williams RL. A Note on Robust Variance Estimation for Cluster-Correlated Data. *Biometrics*. 2000;56(2):645-6.
20. Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations. *Journal of Statistical Software*. 2006;15(2):1-11.
21. Yan J, Fine J. Estimating equations for association structures. *Statistics in Medicine*. 2004;23(6):859-74.
22. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*. 2006;35(1):34.
23. Burton P, WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-78.
24. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. *New England Journal of Medicine*. 2007.
25. Wright S. *Evolution and the Genetics of Populations*: University of Chicago Press; 1968.
26. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, et al. Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet*. 2008;16(12):1413-29.
27. Balding DJ. Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*. 2003;63(3):221-30.
28. Cavalli-Sforza L, Menozzi P, Piazza A. *The history and geography of human genes*. Princeton: Princeton University Press; 1994.
29. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theoretical Population Biology*. 2001;60(3):227-37.

30. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *The Lancet*. 2003;361(9357):598-604.
31. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nature Genetics*. 2004;36(5):512-7.

SNP Spacing	Group	Mean ($\hat{\beta}$)	Rejections of H_0 :
			P<0.05
Every 20 th SNP	Case	1	0.9802
	Control	-0.9999	0.9809
	Neither	-0.0002	0.1138
Every 33 rd SNP	Case	1	0.9864
	Control	-0.9999	0.9858
	Neither	-0.0005	0.0947
Every 100 th SNP	Case	1	0.9886
	Control	-1	0.9891
	Neither	-0.0002	0.0840

Table 1: Results for the linear regression method. One thousand simulations for each SNP spacing are performed generating hypothetical case-control GWAS consisting of 100 individuals in each group. The “Neither” group also consists of 100 individuals. Each simulation run tests each individual within each of the three groups for presence in the hypothetical study. Proportion of rejections of H_0 represents power for individuals in the case and the control groups, and type I error for individuals in neither group.

Analysis	Group	Mean ($\hat{\beta}$)	Reject. H_0 (5% level of sig.)
Linear Regression	Case	1	0.9878
	Control	-0.9999	0.9885
	Neither	0.0004	0.0814
Logistic Regression	Case	5.2266	0.9796
	Control	-5.2262	0.9808
	Neither	0.0022	0.0511
GEE Independence (cluster size = 1)	Case	1	0.9802
	Control	-0.9999	0.9813
	Neither	0.0004	0.0512

Table 2: Results for the linear regression, logistic regression, and GEE independence (with cluster size of 1) approaches for the dataset with a SNP spacing of 100. One thousand simulations are performed generating hypothetical case-control GWAS consisting of 100 individuals in each group. The “Neither” group also consists of 100 individuals. Each simulation run tests each individual within each of the three groups for presence in the hypothetical study for each model. Proportion of rejections of H_0 represents power for individuals in the case and the control groups, and type I error for individuals in neither group.

<i>Analysis</i>	Group	Mean ($\hat{\beta}$)	Reject. H_0 (5% level of sig.)
<i>Linear Regression</i>	Case	1	0.9807
	Control	-0.9999	0.9803
	Neither	0.0009	0.1120
<i>Logistic Regression w/ Offset</i>	Case	5.1539	0.9689
	Control	-5.1530	0.9686
	Neither	0.0049	0.0751
<i>GEE Independence (cluster size = 1)</i>	Case	1	0.9698
	Control	-0.9999	0.9693
	Neither	0.0009	0.0753
<i>GEE AR-1</i>	Case	1	0.9558
	Control	-0.9999	0.9554
	Neither	0.0009	0.0534

Table 3: Results for linear regression, logistic regression, GEE independence (with cluster size of 1) and GEE AR-1 (with a cluster size of 20) on the dataset with a SNP spacing of 20. One thousand simulations are performed generating hypothetical case-control GWAS consisting of 100 individuals in each group. The “Neither” group also consists of 100 individuals. Each simulation run tests each individual within each of the three groups for presence in the hypothetical study for each model. Proportion of rejections of H_0 represents power for individuals in the case and the control groups, and type I error for individuals in neither group.

Regions	Group	Mean ($\hat{\beta}$)	Reject. H_0 (5% level of sig.)
South Vs Central	In Study – South	1.0000	0.9691
	Not in Study – South	0.0052	0.0518
	In Study – Central	-1.0002	0.9664
	Not in Study – Central	-0.0078	0.0557
Central Vs North	In Study – Central	1.0001	0.9735
	Not in Study – Central	0.0129	0.0516
	In Study – North	-0.9999	0.9691
	Not in Study – North	-0.0040	0.0559
South Vs North	In Study – South	1.0002	0.9742
	Not in Study – South	0.0144	0.0495
	In Study – North	-1.0004	0.9718
	Not in Study – North	-0.0152	0.0564

Table 4: Results when cases and controls are drawn from different UK regions in the 1958 Birth Cohort. In each simulation run, 100 individuals from each region are randomly sampled into one arm of a hypothetical case-control GWAS, and another 100 individuals from each region are test individuals under the null hypothesis. The proportion of rejections of H_0 represents power for the individuals in the simulated case-control GWAS, and type I error for individuals not in the study. One hundred simulation runs are performed.

Cohorts	Group	Mean ($\hat{\beta}$)	Reject. H_0 (5% level of sig.)
1958BC Vs NBS	In Study – 1958BC	1.0000	0.9677
	Not in Study – 1958BC	0.0095	0.0539
	In Study – NBS	-1.0006	0.9694
	Not in Study – NBS	-0.0111	0.0573
1958BC Vs CAD	In Study – 1958BC	1.0001	0.9733
	Not in Study – 1958BC	0.0276	0.0549
	In Study – CAD	-1.0006	0.9728
	Not in Study - CAD	-0.0320	0.0571
NBS Vs CAD	In Study – NBS	1.0001	0.9745
	Not in Study – NBS	0.0284	0.0553
	In Study – CAD	-1.0014	0.9717
	Not in Study - CAD	-0.0299	0.0526

Table 5: Results from 100 simulations when 100 cases and 100 controls are drawn from completely different UK cohorts together with 100 individuals from each cohort who are neither cases nor controls. Only the GEE AR-1 model was considered here.

Mean (F_{ST}) based on:		Approach	Power	Type I Error
True: p_{1j}, p_{2j}	Sample: $\widehat{p}_j, \widehat{p}_j^*$			
0.000005	0.005015	Linear Regression	0.9982	0.0627
		GEE Independence	0.9972	0.0489
		GEE AR-1	0.9973	0.0508
0.00005	0.005057	Linear Regression	0.9975	0.062
		GEE Independence	0.9969	0.0497
		GEE AR-1	0.9971	0.0509
0.00025	0.0053	Linear Regression	0.9985	0.0643
		GEE Independence	0.9984	0.0565
		GEE AR-1	0.9982	0.0583
0.0005	0.0055	Linear Regression	0.9987	0.075
		GEE Independence	0.9987	0.0729
		GEE AR-1	0.9986	0.0739
0.001	0.006	Linear Regression	0.9993	0.1224
		GEE Independence	0.9995	0.1383
		GEE AR-1	0.9995	0.1401
0.0049	0.0099	Linear Regression	1	0.7687
		GEE Independence	1	0.9190
		GEE AR-1	1	0.9182
0.0099	0.0148	Linear Regression	1	0.9893
		GEE Independence	1	0.9999
		GEE AR-1	1	0.9999

Table 6: Simulation results testing different values for F_{ST} . Results are based on 100 runs for each scenario,

simulating 5,000 independent SNPs and case-control studies consisting of 100 cases and controls in each. A further 100 individuals from each ancestry are also simulated in each simulation run to derive the measures of type I error. The GEE Independence approach is fitted using a cluster size of 1; the GEE AR-1 approach is fitted using a cluster size of 20.

DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data

Michael Wolfson,¹ Susan E Wallace,^{2,3} Nicholas Masca,⁴ Geoff Rowe,¹ Nuala A Sheehan,⁴ Vincent Ferretti,^{3,5} Philippe LaFlamme,^{3,6} Martin D Tobin,⁴ John Macleod,⁷ Julian Little,^{3,8} Isabel Fortier,^{3,8,9} Bartha M Knoppers^{2,3} and Paul R Burton^{3,4,8,10*}

¹Statistics Canada, Ottawa, Ontario, Canada, ²Centre of Genomics and Policy, Faculty of Medicine, Department of Human Genetics, McGill University, Montreal, Quebec, Canada, ³Public Population Project in Genomics (P3G), Montreal, Quebec, Canada, ⁴Departments of Health Sciences and Genetics, University of Leicester, Leicester, UK, ⁵Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada, ⁶McGill University and Génome Québec Innovation Centre, Montreal, Quebec, Canada, ⁷Department of Social Medicine, University of Bristol, Bristol, UK, ⁸Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, ⁹Department de Médecine Sociale et Préventive, Université de Montréal, Montreal, Quebec, Canada and ¹⁰Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

*Corresponding author. Departments of Health Sciences and Genetics, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, UK. E-mail: pb51@le.ac.uk

Accepted 27 May 2010

Background Contemporary bioscience sometimes demands vast sample sizes and there is often then no choice but to synthesize data across several studies and to undertake an appropriate pooled analysis. This same need is also faced in health-services and socio-economic research. When a pooled analysis *is* required, analytic efficiency and flexibility are often best served by combining the individual-level data from all sources and analysing them as a single large data set. But ethico-legal constraints, including the wording of consent forms and privacy legislation, often prohibit or discourage the sharing of individual-level data, particularly across national or other jurisdictional boundaries. This leads to a fundamental conflict in competing public goods: individual-level analysis is desirable from a scientific perspective, but is prevented by ethico-legal considerations that are entirely valid.

Methods Data aggregation through anonymous summary-statistics from harmonized individual-level databases (DataSHIELD), provides a simple approach to analysing pooled data that circumvents this conflict. This is achieved via parallelized analysis and modern distributed computing and, in one key setting, takes advantage of the properties of the updating algorithm for generalized linear models (GLMs).

Results The conceptual use of DataSHIELD is illustrated in two different settings.

Conclusions As the study of the aetiological architecture of chronic diseases advances to encompass more complex causal pathways—e.g. to include the joint effects of genes, lifestyle and environment—sample size requirements will increase further and the analysis of pooled individual-level data will become ever more important. An aim of this conceptual article is to encourage others to address the challenges and opportunities that DataSHIELD presents, and to explore potential extensions, for example to its use when different data sources hold different data on the same individuals.

Keywords Pooling, analysis, meta-analysis, individual-level, study-level, generalized linear model, GLM, ethico-legal, ELSI, identification, disclosure, distributed computing, bioinformatics, information technology, IT

Introduction

Most known associations between genetic variants and chronic diseases reflect weak effects with typical allelic odds ratios in the range 1.1–1.4.^{1–3} The reliable identification of such effects demands vast data sets.^{1–5} Case–control studies including thousands of cases are required even when interest focuses on the simplest situation: the detection of the direct effects of single nucleotide polymorphism (SNP) variants.^{1–3} Furthermore, when, as is likely, scientific emphasis starts to focus on the study of gene–environment and gene–gene interactions and the exploration of causal pathways more comprehensively, tens of thousands of cases will often be required.¹ Tens of thousands of subjects can also be required to study a quantitative phenotype (e.g. measured blood pressure), because allelic effect sizes may be as small as one-tenth of a standard deviation, or even less.^{6–8}

To achieve sample sizes as large as this, it is often necessary to pool data across multiple studies, and large collaborative consortia have been responsible for much of the recent progress in human population genomics.^{6,8–16} Large-scale data pooling is equally important in other settings too: in mainstream epidemiology¹⁷—particularly in the analysis of formal networks of studies^{18,19}—in public health and health-services research, and in comparative international analysis in the social sciences, including coordinated economic surveillance.^{20,21} Such pooling not only supports the attainment of large sample sizes but can also be used to reduce bias arising from access to a restricted subset of data. But, regardless of its purpose, the sharing of data always raises important ethico-legal issues even when the analysis is mutually agreed. Data privacy, for example, is a hot topic in genomic epidemiology,^{22,23} as well as being a concern for government, industry,^{24,25} the media and even the general public.²⁶ Biomedical science has responded cautiously to these concerns, ensuring that all ethico-legal stipulations are met and that new issues are dealt with carefully, as and when they arise.^{23,27}

Given this caution, it is perhaps surprising that there has been such striking recent progress in detecting genetic associations with complex diseases.^{3,28} In the past three years genome-wide association studies (GWAS)... have reproducibly identified hundreds of associations of common genetic variants with over 80 diseases and traits (<http://www.genome.gov/gwastudies>).⁹ But, in one sense, genomic epidemiology has been fortunate. The class of pooled analysis that has underpinned many of the recent successes,^{6,8–16} just happens to be consistent with the ethico-legal frameworks that large-scale bioclinical studies have had in place over many years. That is, most such studies are permitted to take part in collaborative GWAS based on study-level meta-analysis (SLMA).^{29,30} Here, investigators from each study perform a separate GWAS, and then share the association statistics for each SNP with a designated analysis centre (AC); but the raw data encoding SNP and disease status are *not* shared.^{6,7,11} The AC then performs a meta-analysis to estimate the genetic associations across the consortium as a whole. But, bioscience will inevitably move on from its current focus on simple associations between genetic variants and disease-related traits, to explore causal pathways more thoroughly: e.g. by incorporating gene–environment interactions. This will increase sample size requirements further,¹ making data pooling yet more essential. In addition, data analysis will become increasingly unpredictable and, therefore, exploratory. For example, in a conventional meta-analysis-based GWAS it is clear *a priori* that each study must generate summary statistics to reflect the association of the disease of interest with each of a large number of designated SNPs (e.g. 1 million). This is onerous but it can be pre-specified ahead of time. The required set of summary statistics is far more difficult to predefine if the analysis is to involve gene–environment interactions; environmental and lifestyle factors may be parameterized in many different ways, and identification of the

appropriate parameterization often demands initial exploratory analysis.

Analytic and ethico-legal considerations

Large-scale statistical pooling is typically achieved in one of two ways.^{29,30} First, the individual level data from each of the original data sources can be aggregated to produce one combined data set. This is then analysed as if it were generated by a single study, though study-to-study heterogeneity may necessitate the inclusion of study-specific model terms. This approach may be called individual-level meta-analysis (ILMA). Secondly, appropriate summary statistics can be generated from separate analyses carried out on each independent study, and these then pooled in an SLMA. SLMA is quick and convenient when based on summary statistics that already exist or can be easily derived *de novo*. It is therefore the approach to meta-analysis that is often adopted in public health research, the meta-analysis of randomized controlled trials and, recently, in the pooling of GWAS studies.^{6–8,29–31} But, it has important limitations. First, although it is very convenient to use summary statistics that are already in the public domain, it is important to recognize that they can be biased by selective reporting dependent on findings. In the field of genomic epidemiology this can be particularly problematic.³² Secondly, even when summary statistics are derived *de novo*, SLMA can be restrictive.³⁰ The analysis of all but the simplest of biomedical problems demands a significant element of exploration, but analysis in a conventional SLMA is unavoidably restricted to questions that can be addressed using the particular set of summary statistics that was initially requested.³⁰ If an important new question arises, it can only be answered if the investigators are all prepared to produce the new summary statistics that are required. This can cause serious delays.

In consequence, ILMA would often be preferred to SLMA. But, ILMA raises major ethico-legal challenges. Most notably the sharing of individual level data, sometimes termed ‘microdata’,²⁴ may be prohibited in law. In many jurisdictions, individual-level data are treated as being fundamentally different to aggregate data, and some individual-level data cannot cross certain national boundaries.³³ Even when sharing *is* legal, it may be proscribed by the consents and ethical approvals under which the data were initially collected.³⁴ And, even when—in *principle*—microdata can be shared, that sharing can demand protracted applications for access via scientific oversight committees and ethical review boards.^{35,36} But these barriers are there for a good reason; the relevant ethico-legal considerations reflect important values held by many societies. Individual-level data can disclose identity,²⁴ they may be highly sensitive²⁴ and they may yield unexpected scientific knowledge of great practical or theoretical value,

which the original investigators, funders, national governments and even study participants might feel wary about passing on to a third party.²³ The fundamental importance of these issues is indicated by the fact that they are addressed by the ethico-legal and governance provisions of almost all major bioclinical studies. To illustrate, Box 1 provides exemplar language³⁷ from the ethico-legal documentation of a number of international biobanks and cohort studies, and from the Model Consent Form prepared by the Public Population Project in Genomics (P³G).³⁸ The quotes are not ascribed to particular studies because anonymity was guaranteed as part of the formal agreement under which this ethico-legal documentation was originally shared with P³G.

Resolving a real conflict between ‘competing public goods’

Although ILMA offers many advantages in terms of analytic flexibility,^{29,30} it is therefore clear that ethico-legal restrictions on the transfer of individual-level data to third parties mean that a conventional ILMA approach is often impractical. Since this conflict in ‘competing public goods’ was identified, it has been discussed extensively by the international biobanking community; for example, in forums provided by P³G, Promoting Harmonization of Epidemiological Biobanks in Europe (PHOEBE) and Biobanking and Bio-molecular Resources Research Infrastructure (BBMRI). These discussions have led to the rapid evolution of a novel approach to analysis that could, in theory, circumvent the conflict identified. The proposed approach is named DataSHIELD (Data aggregation through anonymous summary-statistics from harmonized individual level databases). This conceptual article describes the approach proposed, demonstrates that it works in theory, explores its potential uses and extensions, and discusses some of the challenges to be faced in implementing it. It is our hope that by sharing the concept with the broader research community, we will encourage others to work with us in undertaking a pilot implementation.

Methods

The conceptual underpinning of DataSHIELD is straightforward. Modern distributed computing is used to realize the full benefits of ILMA without physically sharing any individual-level data. All data remain on the local computers at their studies of origin and the role of the AC is to coordinate a parallelized analysis of the individual-level data on all of those local computers simultaneously. Critically, the parallelized analysis is so framed that the only information passing back and forth between computers consists of short blocks of computer code specifying

Box 1 Examples of language used in relevant ethico-legal documentation including consent forms and information leaflets

Examples of language used in the ethico-legal documentation of selected international biobanks and cohort studies

(1) Language restricting the scope of data sharing

Use of data restricted to researchers participating in the original study

- (a) ‘All research data are confidential... they will only be used in medical research and [will] remain in the sole use of the participating researchers.’

Use of data restricted to researchers in one country

- (b) ‘Blood and DNA samples may...be distributed to laboratories...around [country] for further research.’
 (c) ‘Research using the anonymous samples will be done by [researchers] ...throughout [country].’

(2) Language ensuring data de-identification

- (a) ‘[Project] will give researchers restricted access to... anonymous samples to conduct [research]...’
 (b) ‘Researchers authorised by [Project] will have access to ... coded information...’
 (c) ‘[Project] researchers or their collaborators at other research institutions... may be allowed access to your DNA sample and medical information, but they will not get... links to your identity.’

Examples of language used in the P³G ‘Model Consent form’

(1) The need to obtain both scientific and ethical approval

- (a) ‘The [Project] gives approved researchers access to data and samples... All researchers will only have access to coded data or samples, in order to protect your privacy. They also have to obtain prior scientific and ethical approval as described above, and their research must fit the purpose of the resource/biobank.’
 (b) ‘The [Project] expects to receive requests and, if approved, provide access to data’.

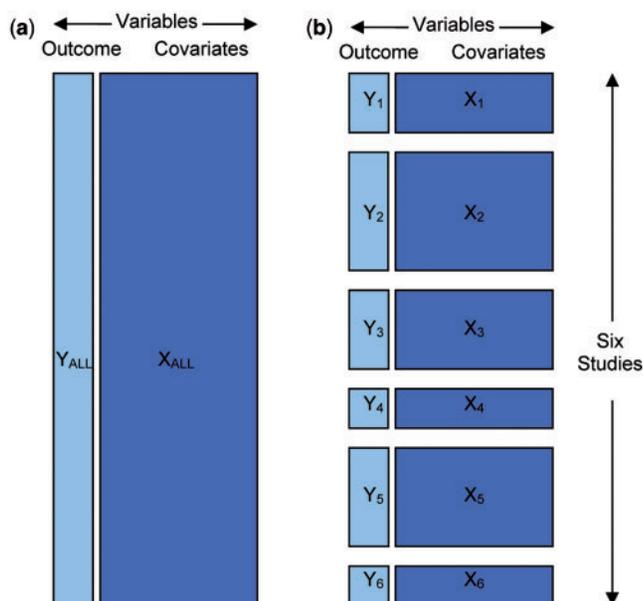


Figure 1 Schematic representation of structure of scientific problems that DataSHIELD is designed to address. (a) One file: all individual-level data pooled together in one large data file. (b) Partitioned: individual-level data held in six separate data files, one for each study

the next analysis required, and low-dimensional summary statistics used in estimating the mathematical parameters of the model (e.g. means or regression coefficients). These items disclose neither the identity, nor the characteristics, of individual study participants.

Figure 1 provides a schematic representation of the class of analytic problems that DataSHIELD is aimed at addressing; here, data are distributed across six sources. The aim is to estimate the statistical parameters that characterize the relationship between an outcome variable Y and one or more explanatory variables X . Here the data are horizontally partitioned:²⁵ i.e. each data set includes all of the variables (X and Y) but on different sets of individuals. A classical ILMA would involve stacking the data matrices from each study to produce one large data matrix (Figure 1a). Under DataSHIELD (Figure 1b), on the other hand, a series of parallel analyses are undertaken simultaneously—using X_j and Y_j in the j th study—and these analyses are synthesized in an appropriate manner to generate estimates pertaining to all six studies simultaneously.

Figure 2 provides a schematic representation of the type of IT infrastructure that might typically be

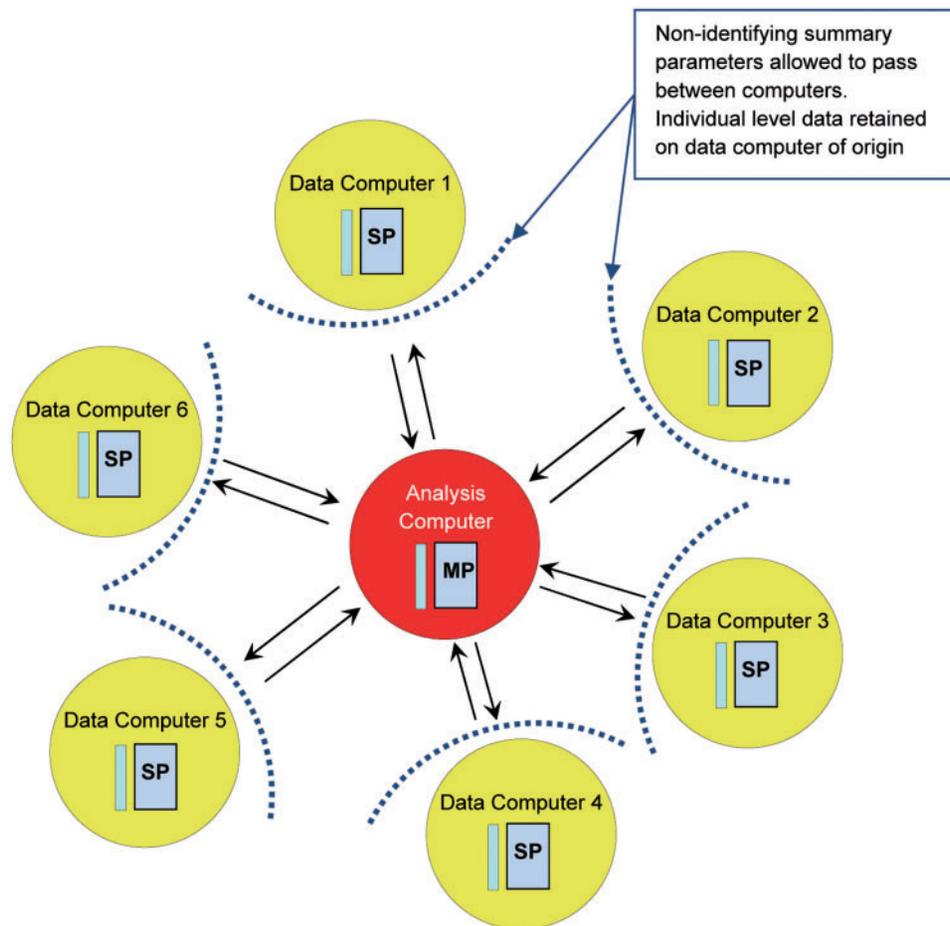


Figure 2 Schematic representation of the structure of DataSHIELD. The computer controlling analysis (heavily shaded circle) is sited at the analysis centre (MP: master process). The data computers (lightly shaded circles) are each sited at one of the study centres involved in the collaborative analysis (SP: slave process). The arrows indicate the flow of analytic instructions and summary statistics. All potentially disclosive individual-level data are secured on the local data computers

required to undertake a DataSHIELD analysis. The computers on which the individual-level data reside at each of the six centres are depicted as lightly shaded circles. One centre is designated the AC and it is a computer (the heavily shaded circle) at that centre that is used to coordinate and execute the analysis. Often, the AC will be one of the studies that are contributing data to the analysis. The analysis software/middleware in DataSHIELD will require two primary components: (i) a master process (MP) that resides on the coordinating computer at the AC; and (ii) a series of slave processes (SPs), each residing on the local data computers. This structure will enable analytic subroutines to be written by the AC, and then transmitted and activated in a suitable software environment (e.g. in 'R'³⁹) on each of the data computers. As an analytic session proceeds, the analysis will evolve and the algorithm that is active on each SP will therefore change. It is the MP at the AC that will control which algorithms are running on which computers at which point in time.

Example 1

Using DataSHIELD to enhance the flexibility of SLMA

Perhaps the simplest application of DataSHIELD might entail the replication of a conventional SLMA. To illustrate this setting, data have been simulated for six hypothetical studies (for details see Supplementary Data: S1 available at *IJE* online) that have assessed peripheral systolic blood pressure (SBP in mmHg^{-1}) as a quantitative outcome variable and two explanatory covariates: AGE (years, centralized by subtracting the mean of 60 years); and an SNP (coded 0, 1 or 2, to reflect the number of copies of a minor allele). An illustrative analysis might involve fitting a multiple linear regression model to estimate a regression intercept ($b_{\text{intercept}}$) and regression coefficients b_{AGE} and b_{SNP} associated with the two covariates. Scientific interest might focus on b_{SNP} to provide an age-adjusted estimate of the increase in SBP associated with each additional copy of the minor allele.

Box 2 Exemplar code and output for Scenario 1**The statistician types:**

```
regression.model<-lm(SBP~AGE+SNP)
results.matrix<-summary(regression.model)$coefficients[,1:2]
```

**Thereby producing a results matrix for each study^a:
for example,**

	Estimate	Std. Error ^b
(Intercept)	125.130	0.2629
AGE	0.203	0.0373
SNP	0.254	0.3907

^aHere, the results shown are for simulated study 6^bStandard Error

If the parallelized analyses are to be undertaken in ‘R’,³⁹ the statistician at the AC might type the two lines of code at the top of Box 2. Using an appropriate scripting language such as Perl⁴⁰ this code could be packaged and transmitted to each of the SPs where it could be piped to R to fit the required regression model on the local data set. This will generate a results matrix (three rows, two columns) comprising an estimate and standard error for each regression coefficient (bottom of Box 2). Additional scripting instructions will then command each study to transmit its results matrix back to the AC. There, the study-specific results can be pooled using an appropriate form of SLMA, to produce parameter estimates and standard errors for all six studies combined. This analysis is detailed in Supplementary Data: S1 (available at *IJE* online).

This DataSHIELD analysis, as outlined, is mathematically equivalent to a conventional SLMA, and all individual-level data remain secure on their computers of origin. But, the first stage (estimation of regression coefficients and standard errors) is controlled remotely by the AC, rather than being carried out by the investigators at each study independently, at the request of the AC. This difference is crucial, because it means that once the initial regression model (Box 2) has been fitted, it is easy to fit a different model that may contain terms for which summary statistics might not, originally, have been requested; for example, one containing an interaction between the AGE and SNP covariates. This would be impossible in a conventional SLMA unless this supplementary analysis had explicitly been pre-specified. This demonstrates that, in principle, DataSHIELD permits SLMA to be undertaken more flexibly. But it offers far more than this. Perhaps most crucially, it allows researchers to make efficient use of an important and versatile class of mathematical models in a manner that is mathematically identical to a full ILMA.

Example 2

Using DataSHIELD to undertake ILMA without sharing the data

Many important analyses in contemporary biopopulation science can be framed as generalized linear models (GLMs).⁴¹ This broad class of models incorporates many forms of regression—e.g. multiple linear regression, logistic regression, Poisson regression and many types of survival analysis. It also subsumes numerous other analytic procedures including *t*-tests, analysis of variance and estimation based on contingency tables.⁴¹ GLMs are usually fitted iteratively using the iteratively reweighted least squares (IRLS) algorithm.⁴² An initial guess at the required regression coefficients is progressively refined, over a number of iterations, until maximum likelihood estimates are obtained. Conveniently, in the present context, updating the coefficient estimates at any given iteration depends solely on an information matrix and a score vector, both of which can be obtained by fitting a single iteration of the same GLM to the individual-level data from each of the collaborating studies one at a time, and by summing them in the AC. The two sums may then be used to update the regression coefficients at that iteration⁴² (for details see Box 3 and Supplementary Data: S2, available at *IJE* online). The regression coefficients and standard errors that are obtained in this manner are *identical* to those that would be obtained by fitting the same GLM to the pooled individual data from all studies combined, but the AC never has access to the individual-level data.

Results

The mathematics underpinning the IRLS algorithm guarantee that the DataSHIELD approach, as implemented in Example 2, will produce the same results as fitting the equivalent GLM to the individual-level data from all studies combined (for details, see Supplementary Data S2 and S4 at *IJE* online). Box 3 provides a concrete example to confirm this claim. It outlines the analysis of a second simulated data set consisting of six hypothetical studies set up to investigate the relationship between the risk of acute myocardial infarction, body mass index (BMI) and an SNP. Full details of the simulation, analysis, computer code and results are provided in Supplementary Data: S3–S6 at *IJE* online). In contrast to the simple model used in Example 1, this GLM incorporates an interaction term to reflect heterogeneity in the magnitude of the increase in risk of myocardial infarction for a given increase in BMI.

As proof of principle, the estimated regression coefficients and standard errors reported at the bottom of Box 3 are precisely the same, rounding error aside, as those derived from a conventional logistic regression model fitted to a single data set comprising the

Box 3 Simulated data example

<p>DATA: six case-control studies</p> <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th>Study</th> <th>Cases</th> <th>Controls</th> <th>Total</th> </tr> </thead> <tbody> <tr><td>1</td><td>962</td><td>1038</td><td>2000</td></tr> <tr><td>2</td><td>1486</td><td>1514</td><td>3000</td></tr> <tr><td>3</td><td>761</td><td>739</td><td>1500</td></tr> <tr><td>4</td><td>142</td><td>158</td><td>300</td></tr> <tr><td>5</td><td>1036</td><td>964</td><td>2000</td></tr> <tr><td>6</td><td>360</td><td>340</td><td>700</td></tr> </tbody> </table>	Study	Cases	Controls	Total	1	962	1038	2000	2	1486	1514	3000	3	761	739	1500	4	142	158	300	5	1036	964	2000	6	360	340	700	<p>MODEL:</p> <p>Outcome variable: CC case control status - cases fulfil formal criteria for an acute myocardial infarction; controls are healthy and population based: 1=case, 0 = control</p> <p>Explanatory covariates: BMI body mass index: [kg/m²] centralised by subtracting 23 kg/m². Increasing BMI is known to have greater health consequences in studies 4, 5 and 6, and the model must therefore allow for study-to-study heterogeneity. This is achieved by adding an interaction term to the model (BMI.456: taking the value 0 in studies 1-3; and the value of BMI in studies 4-6). SNP single nucleotide polymorphism, minor allele frequency = 0.3: 0=no copies of minor allele; 1=one copy; 2=two copies</p> <p>Model formula: $LP = b_{\text{Intercept}} + b_{\text{BMI}} \times \text{BMI} + b_{\text{BMI.456}} \times \text{BMI.456} + b_{\text{SNP}} \times \text{SNP}$</p> <p>CC~binomial(1,exp[LP]/(1+exp[LP]))</p> <p>This is a conventional logistic regression model with an additive genetic effect. An aim of analysis is to derive maximum likelihood estimates for the four regression coefficients: $\hat{b}_{\text{Intercept}}$; \hat{b}_{BMI}; $\hat{b}_{\text{BMI.456}}$; \hat{b}_{SNP}</p>
Study	Cases	Controls	Total																										
1	962	1038	2000																										
2	1486	1514	3000																										
3	761	739	1500																										
4	142	158	300																										
5	1036	964	2000																										
6	360	340	700																										
<p>MODEL FITTING USING DataSHIELD:</p> <p>Data items that are transmitted between computers are highlighted in bold. For additional details see Supplementary Materials. The DataSHIELD analysis is predicated on the assumption that all ethico-legal requirements have been met and that the data are adequately harmonized.</p>																													
<p>Step 1: Analysis Centre (AC) writes code to run one iteration of model in R and transmits this to all six data servers Transmission AC → DS (all 6 DS) short block of computer code (see Supplementary Materials for details).</p> <p>Step 2: AC provides initial guess for four regression coefficients (here, all 0) and passes vector to all six data servers Transmission AC → DS (all 6 DS) vector of regression coefficients: [0, 0, 0, 0]</p> <p>Step 3: Analysis Centre tells each data server to run the computer code using the specified vector of regression coefficients and the data held locally on that server Transmission AC → DS (all 6 DS) instruction to run the control code once</p> <p>Step 4: Control code on each data server generates one matrix and one vector – and these summary statistics are both transmitted to the Analysis Centre. Transmission DS → AC (each DS transmits one matrix and one vector, but each study transmits different values). e.g. in this example, the matrix and vector generated by study 5 and transmitted to the AC are:</p> $(1) = \begin{bmatrix} 500 & 70.56657 & 70.56657 & 297 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 297 & 65.39412 & 65.39412 & 382 \end{bmatrix}; (2) = [36, 487.2951, 487.2951, 149]$ <p>Step 5: AC sums both components across all six studies. The overall sum of the vectors is multiplied by the inverse of the overall sum of the matrices to produce a vector containing four update terms (one for each regression coefficient). These are added to the current vector of regression coefficients to produce refined estimates</p> <p>Step 6: Return to step 2, and transmit the vector of refined estimates of the regression coefficients to all data servers Transmission AC → DS (all DS) vector of regression coefficients: e.g. in this example, the updated vector transmitted at the start of the second iteration is: [-0.32183281, 0.02228647, 0.03911561, 0.53516954]</p> <p>Steps 2-6 are repeated until all estimates stabilize from iteration to iteration. At this point, the refined coefficients passed on in step 2 represent the maximum likelihood estimates of the regression coefficients. Furthermore, standard errors for these coefficients can be obtained by calculating the inverse matrix of the sum of the matrices in step 5, and taking the square root of the elements down the diagonal. In this example the final results obtained are:</p> <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th>Parameter</th> <th>Coefficient</th> <th>Standard Error</th> </tr> </thead> <tbody> <tr> <td>$\hat{b}_{\text{Intercept}}$</td> <td>-0.3296</td> <td>0.02838</td> </tr> <tr> <td>\hat{b}_{BMI}</td> <td>0.02300</td> <td>0.00621</td> </tr> <tr> <td>$\hat{b}_{\text{BMI.456}}$</td> <td>0.04126</td> <td>0.01140</td> </tr> <tr> <td>\hat{b}_{SNP}</td> <td>0.5517</td> <td>0.03295</td> </tr> </tbody> </table>		Parameter	Coefficient	Standard Error	$\hat{b}_{\text{Intercept}}$	-0.3296	0.02838	\hat{b}_{BMI}	0.02300	0.00621	$\hat{b}_{\text{BMI.456}}$	0.04126	0.01140	\hat{b}_{SNP}	0.5517	0.03295													
Parameter	Coefficient	Standard Error																											
$\hat{b}_{\text{Intercept}}$	-0.3296	0.02838																											
\hat{b}_{BMI}	0.02300	0.00621																											
$\hat{b}_{\text{BMI.456}}$	0.04126	0.01140																											
\hat{b}_{SNP}	0.5517	0.03295																											

Box 4 A conventional logistic regression analysis [glm() in 'R'] on pooled data from all six studies combined

	Estimate	SE	z-value	Pr(> z)
Coefficients:				
(Intercept)	-0.32956	0.02838	-11.612	<2e-16
BMI	0.023	0.00621	3.703	0.000213
BMI.456	0.04126	0.0114	3.62	0.000295
SNP	0.55173	0.03295	16.746	<2e-16

individual level data from all six studies combined (Box 4). But (see Box 3 and Supplementary Data: S3–S5 at *IJE* online), information flow between the data sources and the AC is restricted to: (i) repeated instructions from the AC to the data computers to execute each new iteration of the GLM; (ii) non-disclosive summary statistics (one matrix and one vector) passed back from each data computer to the AC at the end of each iteration; (iii) the updated vector of regression coefficients—again non-disclosive—passed from the AC to the data computers at the start of each new iteration. None of these items is disclosive of identity or of sensitive information.

Discussion

This article demonstrates that if all ethico-legal and informatics challenges can be overcome then, in principle, DataSHIELD should enable a full pooled analysis of individual-level data from multiple sources to be undertaken, even when ethico-legal considerations might otherwise obstruct the physical sharing of that individual-level data. At present, DataSHIELD is no more than a concept and there is a quantum leap between proving that the mathematics work and actually implementing the approach in practice. The principal challenges are in developing the IT systems required, in determining whether ethical review committees agree that there is a real problem to be solved and that DataSHIELD provides a workable solution to that problem, and in implementing the local infrastructures at individual biobanks and cohort studies (staff and equipment) to enable its use. These challenges are substantive and it might be argued that publication should await successful implementation. The European Union has recently awarded funding under Framework 7 (the BioSHARE-EU project) to enable preliminary work to develop and pilot the required IT systems and to explore the relevant ethico-legal and social issues. Given that the implementation work will now definitely take place, it is critical to enrol studies, as pilot sites, to work with us in implementing and trialling the method.

Furthermore, the preliminary work will include exploring the fundamental problem with research ethics committees and determining whether they view DataSHIELD as a viable solution. We hope this article will assist studies, biobanks and research ethics committees to determine whether they wish to contribute to such a project.

Development to date has been undertaken by an international group that includes leading bioinformaticians and ethico-legal experts. On the basis of active discourse between these experts and the broader international biobanking community (via P³G and BBMRI), the prevailing viewpoint seems to be that there is a real problem to be overcome and that the fundamental challenges both in the IT and ethico-legal domains can, in principle, be overcome. For example, there is a broad consensus amongst ethico-legal experts that the physical sharing of individual level data between research groups must be subject to appropriate governance and that it is an inescapable fact that the formal documentation and oversight systems in certain studies (see Box 1) proscribes or discourages such sharing. The real challenge, therefore, is to explore whether DataSHIELD provides a workable solution. Bioinformaticians believe that the IT interface should be set up in a manner that actively prevents the AC from tunnelling into the local systems to extract data or other information and/or from fitting models that reveal identifying or sensitive data either directly or by logical deduction. It is therefore commonly argued that the DataSHIELD interface should parse all incoming and outgoing messages and then block and record any request, or series of requests,⁴³ that might, by accident or design, lead to the transmission of inappropriate information. Encouragingly, it seems to be the view of most IT experts that an interface with these characteristics can, in principle, be constructed, and will be feasible to use in practice. This optimistic viewpoint is supported by the fact that secure single-site interfaces already exist allowing external users to specify analyses and then to extract results—but, crucially, no more than results. For example, such an interface is at the heart of the UK's Economic and Social Research Council Secure Data Service.⁴⁴ Provided this optimism proves to be well founded, the majority view amongst ethico-legal and biobanking experts with whom DataSHIELD has been discussed seems to be that DataSHIELD might then be seen as being equivalent to conventional SLMA. This is because, in both settings, information flow between data providers and the AC is restricted entirely to analytic instructions and non-identifying summary statistics. If research ethics committees hold the same viewpoint, any study that is currently unable to contribute to a conventional SLMA-based meta-analysis (including GWASs) should, in principle, be permitted to make use of DataSHIELD, and the formal ethical and governance requirements

should be equivalent. Ultimately, however, the only definitive proof that DataSHIELD will work and will be accepted by ethics review boards is to implement it for real—the publication of this conceptual article is an important step towards that aim.

The mathematics underpinning DataSHIELD is neither novel, nor difficult to implement.^{29,30,41,42} For example, the fitting of a GLM requires no more than a partitioned modification of the conventional IRLS algorithm^{41,42} (see Supplementary Data: S2–S5 at *IJE* online). Rather, the originality of the method lies in the basic concept itself. Interestingly, a similar idea has previously been floated in the technometrics literature,²⁵ and although this means that we cannot claim precedence, it strengthens the academic foundation of the proposal. Critically, the approach seems not to have been noted by statisticians, bioinformaticians or ethicists working in the field of biomedical research and it has neither been promoted nor applied in this important domain. From a technical perspective, our implementation via GLMs might be viewed as a special case of what the technometrics paper refers to as ‘*secure maximum likelihood estimation*’.²⁵ But, the maximum likelihood case is considered only in broad generality in that paper, and there is no specific focus on generalized linear models.²⁵ Furthermore, our implementation via GLMs circumvents some of the ‘complications’ that the technometrics authors note could arise in the more general case.²⁵ Our article therefore brings an exciting and potentially important new concept to the attention of the biomedical research community, and illustrates the practical implementation of that approach via a broad class of models (GLMs) that already has a wide range of applications in bioscience.

The extensive discussion of DataSHIELD since its initial proposal has resulted in a number of important extensions to the concept. The first is to expand the remit of the approach to work with data sets that are vertically²⁵ rather than horizontally partitioned. In contrast to horizontal partitioning (Figure 1), under vertical partitioning the different data sources contain different data items on the same primary set of individuals. Such a scenario occurs commonly when a major cohort study, such as ALSPAC (Avon Longitudinal Study of Parents and Children), links to secondary (often governmental) data sources to enrich the information that are available for analysis.⁴⁵ Critically, the data in such secondary sources are often sensitive and can be protected against misuse by prohibiting their physical release. This same problem arises regularly in cross-jurisdictional analyses being undertaken or overseen by, national statistics agencies such as Statistics Canada or Statistics UK. The mathematics underpinning the solution to the problem of vertical partitioning is ‘*substantially more complex*’²⁵ than that for horizontal partitioning but, in principle, a solution does exist in the form of an approach known as ‘*secure matrix*

products’.²⁵ If this approach can successfully be implemented, this will markedly enhance the utility of the proposed DataSHIELD approach. The second extension that has been proposed is to take advantage of the approach to help bioscience deal with the pooled individual level analysis of data sets that cannot physically be shared, because of their vast physical size. As illustrative examples, such sources may include full genome sequence data or medical images on large numbers of subjects. Finally, we note that DataSHIELD can prove helpful in any meta-analytic setting where analysis at the level of individual patient records would be scientifically desirable, but ethico-legal considerations discourage ILMA. For example, a reviewer has noted that ILMA permits subgroups of subjects in a given study to be added or removed, which might be valuable when exploring the implications of an intention-to-treat analysis. Although care would have to be taken to ensure that such subgroups were not identified in a potentially disclosive manner, DataSHIELD could address this issue if the subgroups were appropriately flagged.

As an important aside, the genomics world is still grappling with the implications of the work of Homer *et al.*²³ A question that is regularly asked of DataSHIELD is whether it would protect against the form of inferential disclosure²⁴ described and explored by Homer *et al.* The simple answer is ‘no’, because disclosure under Homer *et al.* is based on summary statistics reflecting study-wide genotype distributions at each of many SNPs and is therefore totally unrelated to the third party release of individual-level data. This implies that the specific concerns raised by Homer *et al.*²³ cannot be invoked as being part of the rationale for controlling third party release of individual level data and, as a corollary, that these problems cannot be prevented by using DataSHIELD. But, this does raise an obvious follow-up question: ‘Are there *other* circumstances where *summary parameters* can become identifying?’. This is relevant, because DataSHIELD relies on the transmission of summary statistics that are assumed to be non-disclosive. One recognized form of inferential disclosure is termed residual disclosure.⁴³ Here, the differences between a series of closely related summary statistics—that are themselves non-disclosive—permit precise inferences to be drawn about identity and attribute. It is therefore clear that other scenarios do exist in which summary data can become identifying and some of these may be, as yet, unknown. This emphasises the importance of introducing DataSHIELD cautiously. Because the particular set of summary statistics to be transmitted will vary from one class of problem to another, the potential risk of disclosure will require thorough investigation whenever a new class of models is introduced. Some types of model, such as GLMs,^{41,42} are unlikely to be disclosive, not least because they are of low dimension: they typically have few parameters relative to the number of study participants. But the same may not be true of

other models, such as those containing large arrays of random effects.⁴⁶ This latter might restrict the fitting of generalized linear mixed models⁴⁶ (for example by excluding models where there is a random effect for any single subject). On the other hand, it may prove possible to hold the random effects on the local data computers, while transmitting non-disclosive parameters such as the local variance of the random effects. This requires extensive methodological work, but is an area that we believe would be of considerable theoretical interest to many biostatistics research groups.

Regardless of how data pooling is to be approached, two absolute criteria must always be fulfilled. First, all ethico-legal stipulations must be met. This implies that if it is unclear whether the governance rules of a particular study permit DataSHIELD to be used, that uncertainty must be resolved before DataSHIELD is implemented on that study. Secondly, the data to be amalgamated across studies must be sufficiently similar to allow them to be pooled. Two data sets may be said to be harmonized for a given set of variables in a particular scientific setting, if it is valid and feasible to pool them in that setting. DataSHIELD should not be used unless the studies to be pooled are harmonized. This requires a formal judgement to be made, and methods and tools exist to help scientists make this judgement in relation to pre-existing studies: these include the DataSHaPER (<http://www.datashaper.org>) in population genomics and epidemiology, and the methods advocated by the Luxembourg Income Study (<http://www.lisproject.org>) in economics. In addition, it is critical that IT systems are set up so data can be worked on using DataSHIELD.

To finish, we reiterate that our aim in placing DataSHIELD into the public domain at this juncture is to further stimulate active discussion amongst ethico-legal experts, bioscientists, epidemiologists, biostatisticians, health services researchers, social scientists, national statistical offices and IT professionals. It is our hope that interest generated by this article will encourage others to work alongside us in exploring the opportunities presented by this remarkably simple idea. If the key challenges can be identified and met—and there is no reason to believe that they cannot—DataSHIELD can provide an invaluable addition to the growing toolkit (<http://www.P3G.org>) that is facilitating the large-scale pooled analyses that are fundamental to current and future progress in contemporary biomedical and social science.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported as a core element of the research programs of the Public Population Project

in Genomics (P³G) funded by Genome Canada and Genome Quebec, and Promoting Harmonization of Epidemiological Biobanks in Europe (PHOEBE) funded under European Framework 6 (LSHG-CT-2006-518418). The methodological programme at the University of Leicester focusing on genetic statistics and large-scale data harmonization and pooling is also supported by Medical Research Council Project Grant (G0601625), Wellcome Trust Supplementary Grant (086160/Z/08/A), Leverhulme Research Fellowship (RF/9/RFG/2009/0062) and the Leicester Biomedical Research Unit in Cardiovascular Science (National Institute for Health Research). M.W. is Canada Research Chair in Population Health Modelling/Populomics. N.M. is funded by a British Heart Foundation Studentship (FS/06/040), J.L. is a Canada Research Chair in Human Genome Epidemiology.

Conflict of interest: None declared.

References

- Burton PR, Hansell AL, Fortier I *et al.* Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009;**38**:263–73.
- Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. *Nat Protocols* 2007;**2**:2492–501.
- Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;**5**:e1000477.
- Collins FS. The case for a US prospective cohort study of genes and environment. *Nature* 2004;**429**:475–77.
- Khoury MJ. The case for a global human genome epidemiology initiative. *Nat Genet* 2004;**36**:1027–28.
- Newton-Cheh C, Eijgelsheim M, Rice KM *et al.* Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat Genet* 2009;**41**:399–406.
- Newton-Cheh C, Johnson T, Gateva V *et al.* Eight blood pressure loci identified by genomewide association study of 34,433 people of European ancestry. *Nat Genet* 2009;**41**:666–76.
- Repapi E, Sayers I, Wain LV *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2009;**42**:36–44.
- Hindorf LA, Sethupathy P, Junkins HA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 2009;**106**:9362–67.
- Burton PR, Clayton DG, Cardon LR *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* 2007;**39**:1329–37.
- Zeggini E, Weedon MN, Lindgren CM *et al.* Replication of genome-wide association signals in U.K. Samples reveals risk loci for type 2 diabetes. *Science* 2007;**316**:1336–39.
- Frayling TM, Timpson NJ, Weedon MN *et al.* A Common Variant in the *FTO* Gene is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science* 2007;**316**:889–94.

- ¹³ Easton DF, Pooley KA, Dunning AM *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007;**447**:1087–93.
- ¹⁴ Scott LJ, Mohlke KL, Bonnycastle LL *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;**316**:1341–45.
- ¹⁵ Stacey SN, Manolescu A, Sulem P *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2007;**39**:865–69.
- ¹⁶ Saxena R, Voight BF, Lyssenko V *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;**316**:1331–36.
- ¹⁷ Friedreich CM. Methods for pooled analyses of epidemiologic studies. *Epidemiology* 1993;**4**:295–302.
- ¹⁸ Slimani N, Deharveng G, Charrondière RU *et al.* Structure of the standardized computerized 24-h diet recall interview used as reference method in the 22 centers participating in the EPIC project. *Comp Meth Programs Biomed* 1999;**58**:251–66.
- ¹⁹ Harris JR, Willemsen G, Aitlahti T *et al.* Ethical issues and GenomEUtwin. *Twin Res* 2003;**6**:455–63.
- ²⁰ Lynch J, Davey Smith G, Harper S *et al.* Is income inequality a determinant of population health? Part 1. A systematic review? *Milbank Quart* 2004;**82**:5–99.
- ²¹ Backlund E, Rowe G, Lynch J, Wolfson M, Kaplan G, Sorlie P. Income inequality and mortality: a multi-level prospective study of 521, 248 individuals in 50 US States. *Int J Epidemiol* 2007;**36**:590–96.
- ²² Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. *Nat Genet* 2009;**41**:965–67.
- ²³ Homer N, Szlinger S, Redman M *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008;**4**:e1000167.
- ²⁴ Gomatam S, Karr A, Reiter J, Sanil A. Data dissemination and disclosure limitation in world without microdata: a risk-utility framework for remote access analysis servers. *Statistical Science* 2005;**20**:163–77.
- ²⁵ Karr A, Fulp W, Vera F, Young S, Lin X, Reiter J. Secure, privacy-preserving analysis of distributed databases. *Technometrics* 2007;**49**:335–45.
- ²⁶ GCNews. *Health Beats MoD on Equipment Losses*, 2008. <http://www.smarthealthcare.com/equipment-losses> (12 October 2009, date last accessed).
- ²⁷ P3G Consortium, Church G, Heeney C, *et al.* Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet* 2009;**5**:e1000665.
- ²⁸ Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Investigat* 2008;**118**:1590–605.
- ²⁹ Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual- and aggregate-level data. *Stat Med* 2008;**27**:651–69.
- ³⁰ Petitti DB. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. 2nd edn. New York: Oxford University Press, 2000.
- ³¹ Khoury MJ, Little J, Gwinn M, Ioannidis JP. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol* 2007;**36**:439–45.
- ³² Hattersley AT, McCarthy MI. What makes a good genetic association study? *Lancet* 2005;**366**:1315–23.
- ³³ Kaye J. Do we need a uniform regulatory system for biobanks across Europe? *Eur J Hum Genet* 2006;**14**:245–48.
- ³⁴ Zink A, Silman AJ. Ethical and legal constraints on data sharing between countries in multinational epidemiological studies in Europe report from a joint workshop of the European League Against Rheumatism standing committee on epidemiology with the “AutoCure” project. *Ann Rheum Dis* 2008;**67**:1041–43.
- ³⁵ Malfroy M, Llewelyn CA, Johnson T, Williamson LM. Using patient-identifiable data for epidemiological research. *Transf Med* 2004;**14**:275–79.
- ³⁶ Infectious Diseases Society of America. Grinding to a halt: the effects of the increasing regulatory burden on research and quality improvement efforts. *Clin Infectious Dis* 2009;**49**:328–35.
- ³⁷ Wallace S, Lazor S, Knoppers BM. Consent and population genomics: the creation of generic tools. *IRB: Ethics & Human Research* 2009;**31**:15–20.
- ³⁸ Knoppers BM, Fortier I, Legault D, Burton P. The public population project in genomics (P3G): a proof of concept? *Eur J Hum Genet* 2008;**16**:664–65.
- ³⁹ R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2008.
- ⁴⁰ Wall L, Christensen T, Orwant J. *Programming Perl*. 3rd edn. Sebastopol: O’Reilly Media Inc., 2000.
- ⁴¹ McCullagh P, Nelder J. *Generalized Linear Models*. London: Chapman and Hall, 1989.
- ⁴² Aitkin M, Anderson D, Francis B, Hinde J. *Statistical Modelling in GLIM*. Oxford: Clarendon Press, 1989.
- ⁴³ Statistics Netherlands, Statistics Canada, Germany FSO, University of Manchester. *Glossary of Statistical Disclosure Control, Incorporated in Paper Presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. Geneva: UNECE/EUROSTAT, 2005.
- ⁴⁴ ESRC_Secure_Data_Service. <http://www.esrc.ac.uk/ESRCInfoCentre/research/resources/SDS.aspx>. 2009 (21 June 2010, date last accessed).
- ⁴⁵ Ford DV, Jones KH, Verplancke JP *et al.* The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;**9**:157.
- ⁴⁶ Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993;**88**:9–25.