The development and validation of risk assessment tools for non-diabetic hyperglycaemia or undiagnosed diabetes

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Shaun Richard Barber BSc MSc

Department of Health Sciences

University of Leicester

2017

Shaun Richard Barber

The development and validation of risk assessment tools for non-diabetic hyperglycaemia or undiagnosed diabetes

Abstract

Risk assessment tools quantify the risk of an outcome using multiple covariates (risk factors). Risk assessment tools are recommended by diabetes prevention guidelines to allow blood tests to be targeted at individuals with an increased risk of currently having non-diabetic hyperglycaemia or undiagnosed diabetes. This thesis presents work on the identification, development and validation of such risk assessment tools.

Key Findings:

- A systematic review of risk assessment tools for prevalent non-diabetic hyperglycaemia was undertaken. This is the first systematic review to focus on risk assessment tools for prevalent non-diabetic hyperglycaemia. Eighteen risk assessment tools for prevalent non-diabetic hyperglycaemia, and prevalent non-diabetic hyperglycaemia or undiagnosed Type 2 diabetes are summarised.
- An empirical comparison of logistic regression, decision trees, support vector machines and the novel application of chain event graphs for developing risk assessment tools found logistic regression and linear support machine vectors had the best external performance. This is the first empirical comparison for a binary medical outcome in cross-sectional data to include an external validation.
- Risk groups for the Leicester Practice risk score were established, allowing consistent advice to be given across general practices when utilising the tool.
- The Leicester Self-Assessment and Leicester Practice risk scores were externally validated using a nationally representative longitudinal dataset. Both gave comparable performance for identifying prevalent non-diabetic hyperglycaemia or undiagnosed diabetes to the dataset on which they were developed. Furthermore, both identified a small proportion of the population with a substantially increased risk of developing diabetes when utilised in the recommended two-stage screening programme and thus are advocated for use across England.

This thesis aids those wishing to use a risk assessment tool for non-diabetic hyperglycaemia in their selection or development of an appropriate tool, as well as addressing some of the previous limitations of the Leicester Self-Assessment and Leicester Practice risk scores.

Acknowledgements

This thesis was made possible by funding from the University of Leicester department of Health Sciences and Professor Kamlesh Khunti's CLAHRC studentship award.

Firstly, I am especially grateful to Dr Laura Gray for her guidance, encouragement and endless patience throughout the process. My sincere thanks also go to Professor Melanie Davies and Professor Kamlesh Khunti for their supervision, expertise and invaluable support. I feel privileged to have had such a knowledgeable and reassuring group of supervisors.

I am grateful to Raj Gill and Denise Robinson for their hard work in scheduling supervision meetings into busy diaries and helping me to complete a variety of forms.

Many thanks to Professor Jim Smith and Dr Lorna Barclay for their guidance on the chain event graph method and generous gifting of time and support.

I am appreciative to all of my colleagues, both from the Leicester Diabetes Centre and the Department of Health Sciences. There are many individuals who have helped me to endure and often enjoy my PhD project with their kind friendship and reassurance. In particular, I would like to recognise Alison Dunkley, Naina Patel, Andrew Willis, Nikki Perrin, Michelle Hadjiconstantinou, Charlotte Jelleyman and Virag Patel.

I would like to thank my family and friends and all at Christchurch for the support and encouragement. I am extremely grateful to my sisters, my parents and my wife, Eniola, for their unwavering love, faith and understanding.

Table of contents

List of tables

List of figures

List of abbreviations

Abbreviations are written in full the first time they are used in the text of each chapter.

ADA American Diabetes Association

AUROC Area under the Receiver Operator Curve

BMI Body Mass Index

CEG Chain Event Graph

CI Confidence Interval

CRS Cambridge Risk Score

CVD Cardiovascular Disease

ELSA English Longitudinal Study of Aging

EPV Events per Variable

FINDRISC Finnish Diabetes Risk Score

FPG Fasting Plasma Glucose

GTT Glucose Tolerance Test

HbA1c Glycated Haemoglobin A1c

IFG Impaired Fasting Glucose

IGR Impaired Glucose Regulation

IGT Impaired Glucose Tolerance

LAR Least Angle Regression

LASSO Least Absolute Shrinkage and Selection Operator

LPRS Leicester Practice Risk Score

LSA Leicester Self-Assessment

NDH Non-diabetic Hyperglycaemia

NHS National Health Service

NICE National Institute for Health and Care Excellence

NPV Negative Predictive Value

NRI Net Reclassification Improvement

NSC National Screening Committee

OGTT Oral Glucose Tolerance Test

PPV Positive Predictive Value

RAT Risk Assessment Tool

RCT Randomised Controlled Trial

RMSE Root mean square error

ROC Receiver Operator Characteristic

STAR Screening Those at Risk

SVM Support Vector Machine

T2DM Type 2 Diabetes Mellitus

UK United Kingdom

USA United States of America

WHO World Health Organisation

Chapter 1: Introduction

1.1 Chapter outline

This chapter provides the background and general motivation for this thesis. Firstly, Section 1.2 gives an introduction to Type 2 Diabetes Mellitus (T2DM) and non-diabetic hyperglycaemia (NDH), including trends in its prevalence, common co-morbidities and tested interventions. Section 1.3 discusses whether screening for NDH or undiagnosed T2DM is justified in the United Kingdom (UK), considering the advice of the UK National Screening Committee (NSC) and the National Institute for Health and Care Excellence (NICE). Section 1.4: introduces risk assessment tools (RATs), details the outcomes related to T2DM they may detect, the RATs developed for such outcomes in UK populations and discusses the finding of systematic reviews for RATs with such outcomes. Finally, Section 1.5 outlines the aims of this thesis.

Work in this chapter contributed to the following publication:

- Edwardson CL, Gray LJ, Yates T, Barber SR, Khunti K, Davies MJ. Detection and early lifestyle intervention in those at risk of type 2 diabetes. EMJ Reviews. 2014; 2:48-57.

1.2 Type 2 Diabetes Mellitus and non-diabetic hyperglycaemia

Worldwide, the prevalence of diabetes is predicted to rise rapidly over the coming decades, with the 415 million adults with diabetes estimated in 2015 expected to rise to 642 million by 2040 (1). The majority of individuals with diabetes have T2DM, with around 90% of diabetes cases in high income countries thought to be T2DM (2-4). It is estimated almost half (46.5%) of individuals with diabetes are unaware, having not been diagnosed with the condition (1). This is of concern because, as shown in Figure 1.1 below, untreated T2DM can lead to numerous long-term complications including heart disease, blindness and kidney disease (5). T2DM can reduce an individual's life expectancy by as much as 10 years (6). Around 10% of the National Health Service (NHS) expenditure in 2012 was attributed to diabetes, the majority of which is T2DM, and this figure is forecast to rise to 17% by 2035 (7).

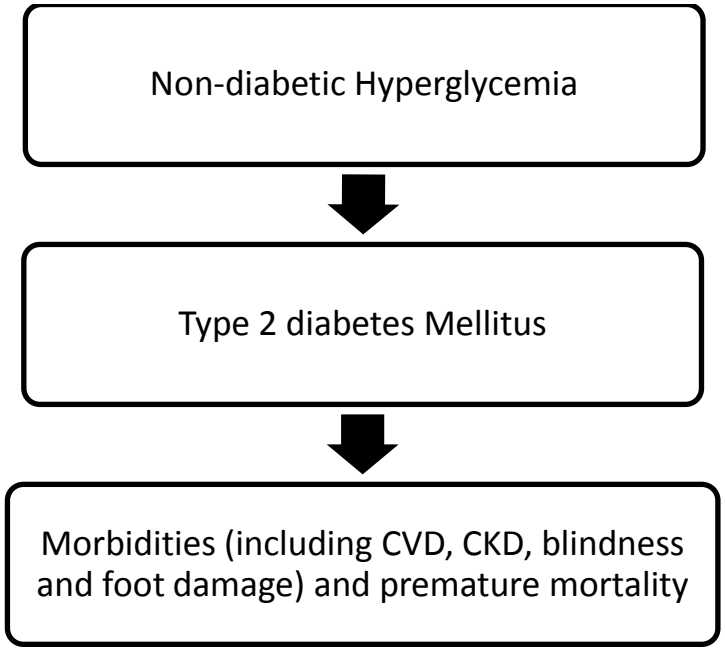


Figure 1.1 T2DM terminology and associated co-morbidities

T2DM is a chronic metabolic disease defined by high blood glucose due to insulin resistance, where the body is unable to use insulin effectively to absorb glucose into its cells, as well as impaired insulin production by the pancreas (8). T2DM can be diagnosed by using either an oral glucose tolerance test (OGTT) or a glycated haemoglobin A1c (HbA1c) test. There are usually two blood measurements taken for an OGTT: fasting plasma glucose (FPG), which is the

glucose measured after 12 hours of fasting, and the 2 hour post-challenge plasma glucose, which is measured two hours after a specified amount of glucose solution is consumed. HbA1c measures the amount of glucose that is attached to the haemoglobin in the blood; this gives a measure of the average level of glucose in the blood over the last eight to 12 weeks, since the red blood cells have a 120 day lifespan (9).

The deterioration of beta-cell function and insulin sensitivity is a gradual process and thus Figure 1.1 is a simplification of the relationship between glucose levels and the development of complications associated with T2DM. The risk of macrovascular and microvascular diseases, which are known to be complications of T2DM, increase as glucose levels increase (10,11). The risk of morbidities and mortality increase exponentially as glucose levels increase and begin before the glucose levels which define T2DM have been reached (11). The cut-off points for T2DM defined by the World Health Organisation (WHO) which are widely accepted are as follows (12,13):

- FPG ≥7.0mmol/l
- 2-h plasma glucose ≥11.1mmol/l
- HbA1c ≥6.5% (≥ 47mmol/mol)

All three cut-off points are based on the increased levels of diabetes related complications observed in each of the measures; the HbA1c cut-off is based on studies which show moderate retinopathy rises above ‘background levels’ at this point (14).

Several studies have demonstrated that interventions to reduce glucose, blood pressure and cholesterol in T2DM patients can decrease the risk of complications (15-18). However individuals with T2DM can remain undiagnosed for several years before the condition is detected in clinical practice, with a recent study using data from Italy predicting on average individuals live with the condition undiagnosed for between four and six years (19). This results in treatment to reduce hyperglycaemia and cardiovascular disease (CVD) risk factors being delayed (20). The International Diabetes Federation advises that if individuals are diagnosed earlier with T2DM then they can start to manage the disease and thus have a greater chance of avoiding complications (21). Despite

this advice, there is no firm evidence that earlier detection of T2DM reduces the risk of complications (22). The ADDITION trial was a randomised controlled trial (RCT) of individuals with screen diagnosed T2DM comparing intensive treatment to conventional treatment, no significant difference was found in the incidence of cardiovascular events and death (23). Although, lower than predicted rates of complications for both the intensive treatment and conventional treatment arms were observed, as this study did not have an unscreened arm for comparison it does not provide direct evidence that earlier detection of T2DM leads to a reduction in complications (22). Simmons et al. present evidence that mortality rates (all-cause, cardiovascular or diabetes-related) over 10 years were not reduced by diabetes screening at the population level (24).

NDH is a condition where an individual has glucose raised above normal levels but not in the T2DM range (25). As well as an increased risk of microvascular disease being observed for individuals with NDH (10); individuals with NDH also have a higher chance of developing T2DM than an individual with normal glucose levels (26,27). One specific type of NDH is impaired glucose regulation (IGR), this is where either impaired glucose tolerance (IGT) or impaired fasting glucose (IFG) has been identified using an OGTT (4). Estimates of progression to T2DM within a year suggest those with isolated IGT have over five times the risk, those with isolated IFG have 7 times the risk and those with both IGT and IFG have over 12 times the risk compared to normoglycemic individuals (10). IGT is defined as a 2-h OGT above 7.8mmol/l but below 11.1mmol/l, while IFG is defined as FPG above 6.1mmol/l but below 7mmol/l. It was estimated in 2013 that 6.7% of 20-79 year olds in the world had IGT, with this figure projected to rise to 7.8% by 2040 (1).

It is also recommended that HbA1c levels raised above normal levels but not elevated enough for a diagnosis of T2DM should be classified as NDH (6). With follow-up studies having shown similar rates of progression to diabetes from HbA1c defined NDH as seen for IFG (10). There is no agreed consensus on the HbA1c range that should be classified as at high risk of diabetes, with the International Expert Committee and the UK-based NICE recommending it be

6.0- 6.4% (42– 46mmol/mol); whereas the American Diabetes Association (ADA) suggests 5.7 -6.4% (39 –46 mmol/mol) (14,28,29).

The term to use for the condition NDH has been hotly debated (30). 'Pre-diabetes' which was once commonly used, has been criticised for being misleading by some experts as it implies that progression to T2DM is inevitable, which is untrue (14). Some critics of the 'Pre-diabetes' term prefer the phrase 'at high risk of diabetes' (30). However this term can be confusing as it does not specify that high risk has been defined by a blood test measurement rather than a statistical model. Therefore, throughout this thesis I use the term NDH for the condition as it avoids these two issues. Where necessary, the exact type of NDH, for example IGR or by HbA1c, will be specified.

Lifestyle and pharmacological interventions have been shown to delay or prevent T2DM in those with NDH, with several studies showing that the risk of progression could be reduced by between 30-60% (31-33). Furthermore, studies have shown good evidence that the benefits of such lifestyle interventions can be long lasting, with one study still seeing participants in the intervention arm enjoying a reduced risk of progression to T2DM 14 years after the intervention had ended (34). A 47% reduction in severe retinopathy, which was statistically significant at the 5% level, was also found for individuals in the intervention arm of this study (35). Yet many individuals have NDH without knowing it and thus the condition may progress untreated along the pathology for many years (21).

The recently launched NHS Diabetes Prevention Programme (DPP) aims to identify individuals at high risk of developing T2DM and refer them to an evidence-based behaviour change programme in order to aid them to reduce their risk (36). By 2020 the programme will be running across the whole country with a projected 100,000 individuals being referred each year (36). In order to identify individuals with NDH or undiagnosed T2DM early and offer these interventions, screening of individuals is required. The next section, 1.3, will discuss guidelines from the UK NSC and NICE on when to screen for a condition, in particular their guidance on screening for NDH and undiagnosed T2DM.

1.3 Should screening for non-diabetic hyperglycaemia or undiagnosed Type 2 diabetes mellitus take place?

The UK NSC guidelines set numerous criteria that should preferably be met before they will recommend screening for a condition (37). Firstly, the condition should be a vital and well-understood health problem for which all feasible cost-effective primary preventions have been employed. There needs to be a validated test, which is acceptable to the population, with agreed cut-off levels and policy on the action to be taken after a positive test. Evidence of an effective treatment or intervention is required; furthermore health-care providers should have implemented agreed evidence-based policies in order to best manage patient outcomes. RCTs should have shown that the whole screening programme decreases mortality and morbidity. Evidence is also necessary to show that the screening programme is: acceptable to both healthcare professionals and the public, is the most cost-effective option available and its benefits are greater than any damage it causes. Before a screening programme begins, sufficient staffing and facilities need to be obtainable. Finally there should be scientific rational for the eligibility criteria.

A short report by NICE on screening for T2DM highlights that some of the criteria for screening are not met in the case of T2DM; furthermore both this report and NICE public health guidance 38 do not advocate universal screening (28,38). Instead both UK NSC and NICE endorse selective screening for T2DM, advising screening takes place in stages with the first stage being to identify those with an increased risk of having the condition using non-invasive RATs. Under this staged approach, those with high risk scores should then receive a blood test, either FPG or an HbA1c. Finally those with high measurements from the blood test, that have not displayed any symptoms of diabetes, should have a second confirmatory blood test to either confirm T2DM or NDH or to rule out the result of the first blood test as abnormally high. Both UK NSC and NICE recognise that in addition to identifying individuals with T2DM earlier than waiting for symptoms to lead to a diagnosis, screening for T2DM will also identify individuals with NDH who will benefit from proven lifestyle change programmes (28,38).

Whether to use OGTTs or HbA1c tests of blood glucose for screening and diagnostic tests has been widely debated, this is because the two find overlapping but different groups of individuals for both NDH and T2DM (39). The UK NSC report puts forward using HbA1c for the screening blood test followed by either an OGTT or HbA1c for the diagnostic test, stating there is growing evidence to warrant HbA1c's use (38). Additionally suggesting that uptake of OGTT would be lower due to the inconvenient and time-consuming nature of having to fast overnight and then have a two hour test, as well as raising concerns that OGTTs measurements can vary from week to week. The NSC report recommends that individuals with HbA1c ≥6.0% on the screening test are followed up with a diagnostic test, though it warns with this cut-off point 20% of individuals with OGTT defined T2DM will be missed (38).

NICE public health guidance 38, displayed in Figure 1.2, recommends individuals found to have undiagnosed T2DM receive standard care for newly diagnosed T2DM; while it advises those found to have NDH are offered group and individual level quality–assured intensive lifestyle interventions (28). In addition to these interventions it states individuals with NDH should be offered a blood test at least once a year, with individuals with high risk scores but normal glucose levels being reassessed at least every three years.

From the above discussion it is clear that RATs are a key element of the screening programme; they are the focus of this thesis and will therefore be introduced and explained in the next section, 1.4.

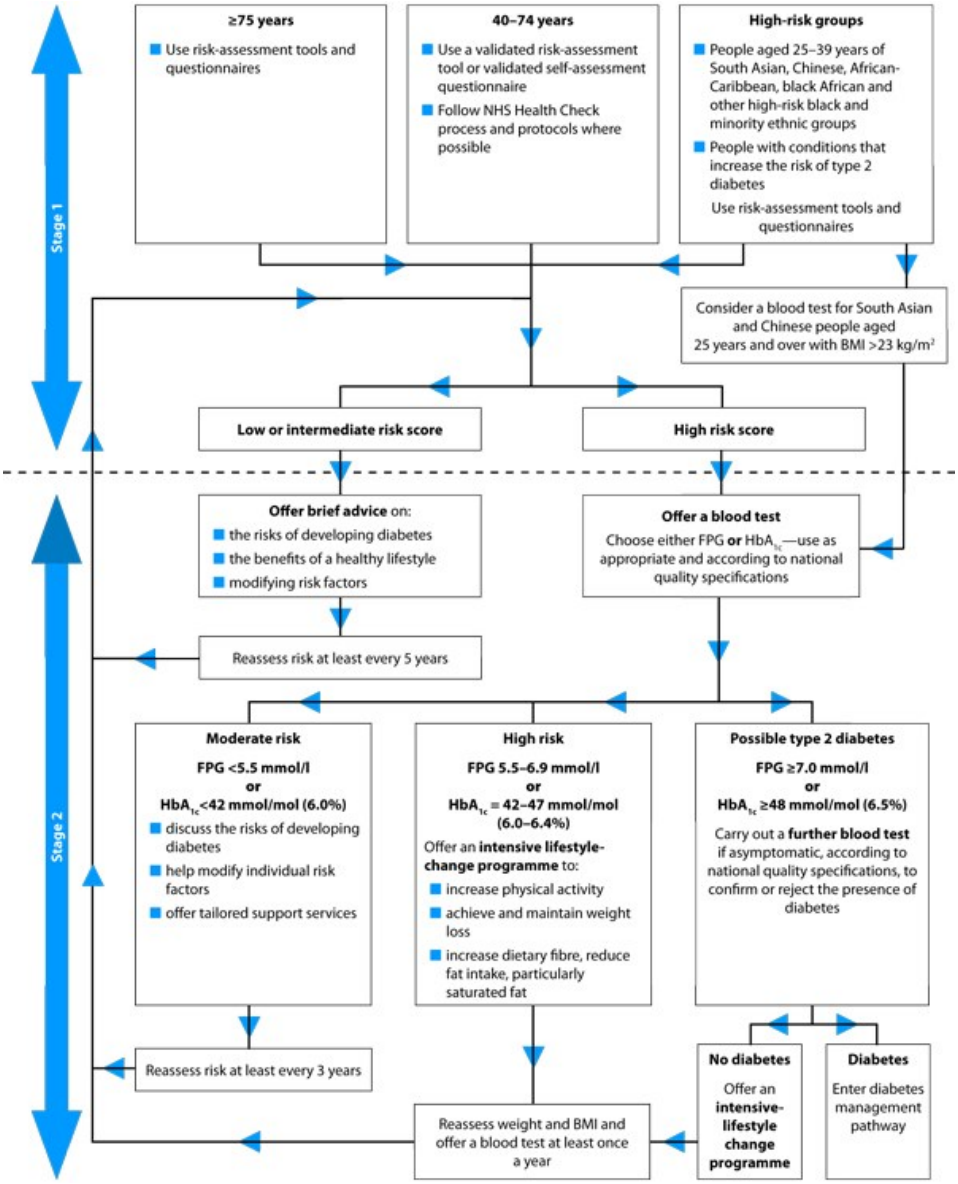


Figure 1.2 NICE guidelines on identifying and managing risk of T2DM (28)

1.4 Risk assessment tools

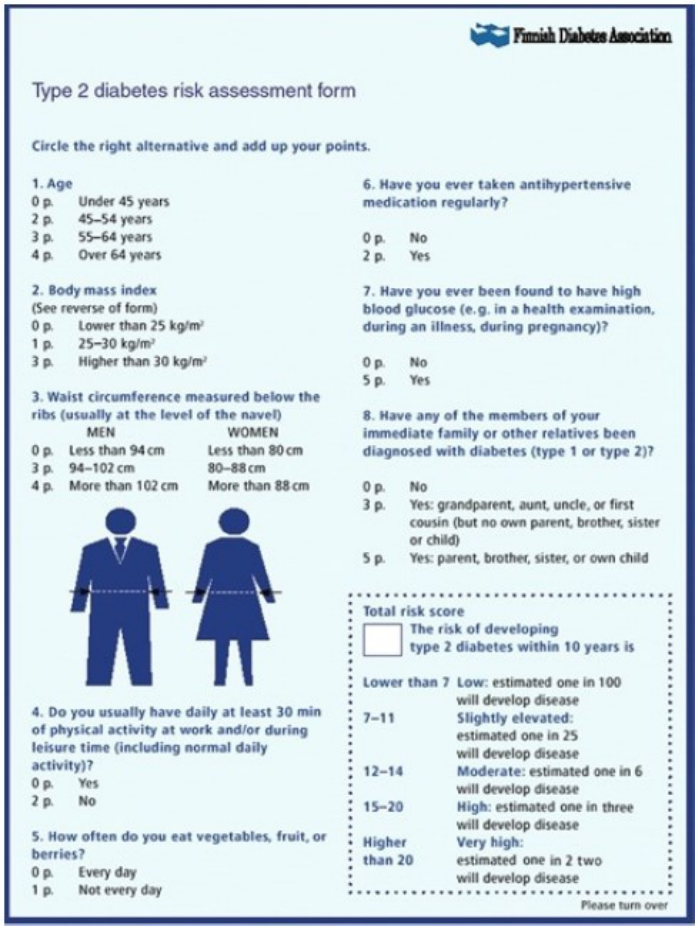
RATs help to optimise the often limited resources required for detecting diseases by allowing screening to be targeted at those with the highest risk. In addition to this they can make invasive tests, such as blood tests, more acceptable to individuals as the chances of having the disease is higher than if universal screening was being carried out. RATs can be developed to attribute a risk score, based on the probability of having a particular health outcome, for an individual given their characteristics (risk factors); the higher the risk scores, the greater the probability of an individual having a particular outcome. These characteristics could be variables that are already stored in a database or factors that are cheap and easy to collect from individuals, for example through a self-completed questionnaire. RATs have the additional advantage that they educate individuals about their risk factors informing them of facts that using a blood test alone would not.

Two key aspects of RATs' accuracy of performance are their discriminative ability and calibration (40). Discrimination informs how well a RAT splits individuals into those who have the outcome of interest and those who do not. The area under the receiver operator curve (AUROC) or C-statistic is commonly used to measure the discrimination of a RAT. The AUROC is detailed in section 2.3.2 of this thesis, briefly the AUROC gives the probability that a randomly chosen case will have a higher risk score than a randomly chosen non-case. An AUROC of 0.5 indicates that the RAT discriminates no better than chance alone; with discrimination of above 0.70 deemed to be good and one above 0.8 excellent (41). Calibration refers to whether the predicted probabilities of having the outcome match the observed probabilities of having the outcome across the whole RAT. There are several 'goodness-of-fit' statistics which can be calculated such as the Hosmer-Lemeshow, Brier score or Chi-squared test value; plots of predicted probabilities against actual probabilities for grouped intervals of a risk score can also be used to visually assess calibration. Thresholds can be chosen for risk scores, with individuals scoring above a chosen threshold being considered to have screened positive by the risk score. Statistical metrics for the performance of binary screening decisions, such as

those produced by using a threshold for a risk score are detailed in section 2.3.1 of this thesis.

The Finnish diabetes risk score (FINDRISC) was one of the first scores to be developed for identifying those at risk of developing diabetes over the next 10 years (42). Baseline characteristics were collected through participants completing questionnaires and attending non-invasive clinical examinations, such as weight measurement, in 1987 (for development data) and in 1992 (for validation data). Participants were then followed-up for the outcome of drug-treated diabetes until the end of 1997. A logistic regression model was fitted for this outcome, with all independent variables split into categories. The independent variables considered were easy to obtain values that did not require laboratory tests or any specialist skill to measure.

Figure 1.3 shows the variables included in FINDRISC; age, body mass index (BMI), waist circumference, antihypertensive drug therapy and history of high blood glucose were all independent significant predictors; while physical activity and vegetable or fruit consumption which were not significant predictors were included to emphasise their importance in diabetes prevention. A score value was assigned to each category of each variable based on its coefficient in the logistic regression model, with the FINDRISC score being the sum of the scores for the variable categories an individual has. The score developed discriminated excellently with an AUROC of 0.85 and 0.87 for the development and validation cohorts respectively. However the original paper does not report any calibration measurements.



Finnish Diabetes Association

Type 2 diabetes risk assessment form

Circle the right alternative and add up your points.

1. Age
0 p. Under 45 years
2 p. 45–54 years
3 p. 55–64 years
4 p. Over 64 years

2. Body mass index
(See reverse of form)
0 p. Lower than 25 kg/m²
1 p. 25–30 kg/m²
3 p. Higher than 30 kg/m²

3. Waist circumference measured below the ribs (usually at the level of the navel)

	MEN	WOMEN
0 p.	Less than 94 cm	Less than 80 cm
3 p.	94–102 cm	80–88 cm
4 p.	More than 102 cm	More than 88 cm

4. Do you usually have daily at least 30 min of physical activity at work and/or during leisure time (including normal daily activity)?
0 p. Yes
2 p. No

5. How often do you eat vegetables, fruit, or berries?
0 p. Every day
1 p. Not every day

6. Have you ever taken antihypertensive medication regularly?
0 p. No
2 p. Yes

7. Have you ever been found to have high blood glucose (e.g. in a health examination, during an illness, during pregnancy)?
0 p. No
5 p. Yes

8. Have any of the members of your immediate family or other relatives been diagnosed with diabetes (type 1 or type 2)?
0 p. No
3 p. Yes: grandparent, aunt, uncle, or first cousin (but no own parent, brother, sister or child)
5 p. Yes: parent, brother, sister, or own child

Total risk score
The risk of developing type 2 diabetes within 10 years is

Lower than 7	Low: estimated one in 100 will develop disease
7–11	Slightly elevated: estimated one in 25 will develop disease
12–14	Moderate: estimated one in 6 will develop disease
15–20	High: estimated one in three will develop disease
Higher than 20	Very high: estimated one in 2 two will develop disease

Please turn over

Figure 1.3 FINDRISC Questionnaire (43)

1.4.1 Outcomes of risk assessment tools relating to T2DM

RATs can be developed for different types of outcomes by using different types of datasets. Cross-sectional studies collect data at one specific time-point and therefore give the prevalence of a disease; thus if a cross-sectional dataset is collected for a population of individuals without a diagnosis for a particular disease it can be utilised to develop a RAT for detecting the risk of having that disease without being diagnosed, for example undiagnosed T2DM. Such RATs are often referred to as diagnostic models (44). On the other hand, prospective cohort studies collect follow-up data, either the time-to-event (diagnosis with a disease or first adverse event) or simply whether an outcome (diagnosis of disease or occurrence of an adverse event) has occurred by one specific-point in the future. The latter can be used to develop a RAT to identify those that are most likely to progress to a particular outcome by a certain point in the future, as was the case for FINDRISC, this outcome is called the cumulative incidence. Whereas using time-to-event data allows a RAT to be created for the incidence rate of an event or disease, this RAT will identity those which are more likely to have an outcome at any point in time. RATs that specify the risk of a future event are frequently referred to as prognostic models (44).

RATs for the following outcomes are useful in either identifying individuals with undiagnosed T2DM or those likely to develop T2DM in the future and hence as stated in section 1.3 can be useful in the screening programme:

- Prevalent NDH and undiagnosed T2DM
- Prevalent undiagnosed T2DM
- Incidence (either incidence rate or cumulative incidence) of T2DM

RATs are not only useful in identifying those likely to be in or progress to the first two stages of Figure 1.1, they are also valuable for assessing the risk of other outcomes along the T2DM pathology, for example they could be utilised to identify which individuals with T2DM are at the biggest risk of developing chronic kidney disease over the next five years.

1.4.2 UK risk assessment tools

In the UK four RATs have been developed in the area of T2DM (45-48). Firstly the Cambridge Risk Score (CRS) was developed using cross-sectional data on 549 individuals without diagnosed diabetes from Ely general practices along with data on 101 cases of newly diagnosed T2DM in the last year from general practices in Wessex (48). Cross-sectional data on 528 individuals from the Ely general practice were used as 'test data' to assess the model. A logistic regression model was built for the outcome of T2DM (undiagnosed or newly diagnosed); independent variables considered were all routinely collected data. Figure 1.4 below shows the variables that were significant and thus included in this RAT. The CRS had excellent discrimination in the test group with a C-statistic of 0.80.

	Risk score	Characteristic
α	−6.322	Constant
$\beta_1 x_1$	−0.879	Female
$\beta_2 x_2$	1.222	Prescribed antihypertensive medication
$\beta_3 x_3$	2.191	Prescribed steroids
$\beta_4 x_4$	0.063	*x* age in years
$\beta_5 x_5$	0	Body mass index <25
	0.699	Body mass index =25 to 27.49
	1.970	Body mass index =27.5 to 29.99
	2.518	Body mass index ⩾30
$\beta_6 x_6$	0	No first degree relative had diabetes
	0.728	Parent or sibling had diabetes
	0.753	Parent and sibling had diabetes
$\beta_7 x_7$	0	Non-smoker
	−0.218	Ex-smoker
	0.855	Current smoker

[a]Probability of having Type 2 diabetes $= \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2 \cdots\cdots+\beta_n x_n)}}$

Figure 1.4 Cambridge Diabetes Risk Score (48)

The CRS was developed using a notational sample which may have detrimentally affected the parameter estimates, this was done due to a lack of cases. The dataset consisted of predominantly Caucasians and the RAT does not include ethnicity as a risk factor so will under estimate risk in black and

minority ethnic (BME) groups (49). A study testing the performance of the CRS in Caribbean and South Asian individuals living in the UK, found that further studies were required to establish ethnicity specific cut-points (50).

The QDScore was developed using over 16 million person-years of observations from an ethnically and socio-economically diverse prospective cohort drawn from 355 general practices around England and Wales (47). It included independent variables which individuals would be likely to know themselves or variables that are readily available in their patient records. A Cox proportional hazards model was used to produce separate a risk equation for men and women for the outcome of incident diabetes recorded in patients records over the ten year follow-up period. The RATs includes the following variables: self-assigned ethnicity, age, sex, body mass index, smoking status, family history of diabetes, Townsend deprivation score, treated hypertension, cardiovascular disease, and current use of corticosteroids (47). Fractional polynomial terms were included in the model for age and BMI; as well as interactions between the age terms and each of smoking status, family history of diabetes and the BMI terms. Due to the use of the Townsend deprivation score, interaction terms and fractional polynomial terms the QDScore cannot be paper based. However the RAT is available on a simple web calculator meaning, it can be easily used by both clinicians and the general public. The external validation of this RAT, using over 1.2 million person-years of observations, showed both good discrimination and calibration. The QDScore was designed to identify those likely to develop T2DM over the next ten years and hence is suitable to be used to identify those whom require intervention or lifestyle change rather than those that currently require a T2DM test.

Two risk scores have been developed for use in the multi-ethnic UK setting (45,51) using ADDITION-Leicester, a dataset containing 6,390 individuals aged 40- 75 years old which was yielded by invitation of a population based sample of individuals without diagnosed diabetes from Leicester and the surrounding county for screening between 2004 and 2008. Participants completed a biomedical questionnaire and had anthropometric measurements taken. All participants that were screened received an OGTT using 75g of glucose as well as having their HbA1c measured. Both the risk scores were externally validated

using over 3,000 individuals screened in the Screening Those At Risk (STAR) study (45,51) which was carried out in Leicestershire.

The first of the scores developed in Leicester, the Leicester Self-Assessment (LSA) score (51), was developed with the aim that it could be completed by a lay person to identity those with NDH or undiagnosed T2DM. This aim was met by only including variables individuals would know about themselves already or could easily measure, for example waist circumference; and by assigning a whole number score to each category of each variable based on the logistic regression model coefficient so that individuals can calculate their score by adding whole numbers. The variables included in the RAT, shown in Figure 1.5, were all significant apart from sex which was included due to previous evidence of its predictive ability for this outcome. The discrimination of this risk score was good with an AUROC for the external validation data of 0.72. Figure 1.5 also displays the risk groups used to communicate the meaning of a particular LSA score and advise individuals what action to take based on their LSA score.

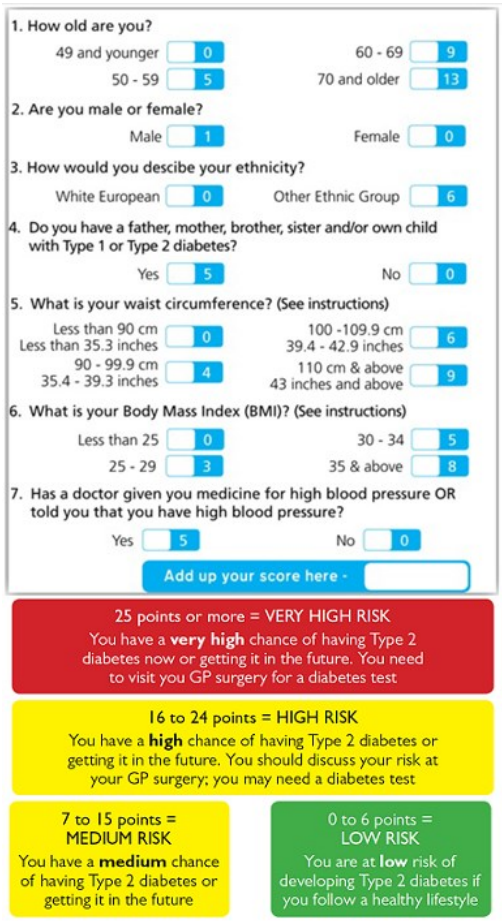


Figure 1.5 Leicester Self-Assessment Risk Score and its risk groups (44)

The second score developed in Leicester, the Leicester Practice Risk Score (LPRS) (45), was developed in order for general practices to utilise data which are routinely stored in primary care to identity individuals most likely to have NDH or undiagnosed T2DM. It is also the first score developed which includes HbA1c to define T2DM in its outcome, which is helpful given the new guidelines (12) and the fact that more practices are using this as the test for T2DM. This score was also developed using a logistic regression model. As the score was designed to be calculated using software the continuous variables were kept continuous. Figure 1.6 details the equation for calculating the LPRS given in the original publication. The discrimination was good with AUROCs in an external dataset of 0.71 for undiagnosed T2DM by OGTT, 0.69 for NDH or undiagnosed T2DM by OGTT, 0.69 for HbA1c ≥6.5% and 0.67 for HbA1c ≥6.0% (45).

LPRS = 0.0408359 X age (years)

+ 0.1839942 (if male, no change if female)

+ 0.7565977 (if Black, Asian or of minority ethnicity, no change if White European)

+ 0.08206 X BMI (kg/m^2)

+ 0.4770517 (if family history of diabetes, no change otherwise)

+ 0.5498978 (if on antihypertensive medication, no change otherwise)

Figure 1.6 Leicester Practice Risk Score (45)

Implementation of LPRS in two prevention trials has shown that it produces a significantly higher yield of undiagnosed T2DM and NDH than population screening programmes in the same area, showing it is an inexpensive way to target screening at individuals with the greatest risk (52). Interestingly, uptake of blood tests for those identified as high risk was similar to that seen for population-level screening. However, this has not been the case for other studies which have shown risk stratification to increase the attendance at screening (53), and the low uptake in this case could well be due to individuals not wanting to participate in the prevention trial or the low level of risk communication.

1.4.3 Discussion of the finding of systematic reviews into risk assessment tools with outcomes of NDH and/or T2DM

Several systematic reviews into RATs for the outcomes listed in section 1.4.1 have been published (54-59). Table 1.1 lists these systematic reviews along with the outcome(s) of the RATs included in the review, the different methods used to build the RATs included and the number of RATs found. The systematic reviews reveal a vast number of RATs have been developed for both the outcomes of prevalent undiagnosed T2DM and incident T2DM. However, only three RATs with an outcome of prevalent NDH and undiagnosed T2DM have been included in these reviews. This is due to the fact that although two of the reviews included such RATs, their search terms focused on finding RATs with either prevalent undiagnosed T2DM or incident T2DM rather than NDH (57,59).

Table 1.1 Systematic reviews for NDH and/or T2DM outcomes

First Author of review	Outcome(s) of risk assessment tools in review	Number of risk assessment tools included in review	Methods used to develop risk assessment tools found
Abbasi (54)	Incident T2DM	25	Logistic regression model (76%), Cox survival model (12%) & Weibull Survival model (12%)
Brown (55)	Prevalent undiagnosed T2DM[a]	17	Logistic regression model (88%) & Decision tree (12%) [RAT with incident T2DM as outcome used Cox survival model]
Buijsse (56)	Incident T2DM	>46 (as 46 studies included, some with multiple RATs)	Logistic regression, Cox survival model (method not stated for individual RATs)
Collins (57)	Prevalent undiagnosed T2DM and NDH, prevalent undiagnosed T2DM or incident T2DM[b]	43	Logistic regression model (72%), Cox survival model (19%), Decision tree (7%) & Weibull survival model (2%)
Noble (58)	Incident T2DM	94	Not stated in review[d]
Thoopputra (59)	Prevalent undiagnosed T2DM and NDH, prevalent undiagnosed T2DM or incident T2DM[c]	41	Logistic regression model (74%), Cox survival model (17%) & Decision tree (9%)

[a] Also included one RAT for incident T2DM
[b] Also included three RATs with outcome of *prevalent NDH* and undiagnosed T2DM, three RATs with outcome of *prevalent undiagnosed and diagnosed T2DM*
[c] *Only includes one RAT* with outcome of *prevalent at NDH* and undiagnosed T2DM
[d] *Through searching available papers the majority were developed using Logistic regression or Cox survival model with one developed using Weibull survival model*

This section will evaluate the findings of these reviews in light of guidance for research into RATs which are intended to be used in clinical practice. There are three important stages of research for RATs which are intended to be used in practice: model development with internal validation, external validation and impact studies of RAT's used in clinical practice (40).

The systematic reviews find that over 100 RATs for T2DM outcomes have been developed and internally validated. As displayed in Table 1.1, RATs were developed using four different methods. The technique used depended on whether the outcome was a prevalent or incident outcome. As can be seen in Table 1.1, the majority of RATs for incident T2DM were developed using logistic regression, several used a Cox survival model and a handful were derived from Weibull survival models. Logistic regression was also the most common method for RATs for the prevalent outcomes, while the rest used decision trees (classification trees).

RATs based on Logistic, Cox or Weibull regression calculate an individual's risk score using their values for the selected variables along with the model's beta-coefficients. Normally a score is allocated for each variable by multiplying the value for each variable by its beta-coefficient given in the regression model, although these beta-coefficients are frequently rounded. Adding the score for each variable gives the overall risk score of an individual. Although some risk scores were yielded using a more crude method of assigning beta-coefficients into score points, with their points being neither proportional to the beta-coefficients or a rounded version of the beta-coefficients (42). Logistic regression has a binary outcome, hence RATs based on Logistic regression for incident T2DM were for cumulative incidence. It has been suggested by some that survival models better represent the prospective nature of the data (56).

Decision trees separate individuals into different groups using a series of stages. At each stage the chosen factor (continuous or categorical) is split into two categories, choosing the variable and cut-point that best separate individuals into those with a high chance of having the outcome and those with a high chance of not having the outcome (60). This process of splitting the individuals into different groups is repeated until the discriminative benefit of adding more splits is smaller than a predefined penalty. Decision trees are prone to being unstable unless statistical techniques, such as bootstrapping, boosting or bagging are applied (61); which develop the decision tree using several resampled versions of the data.

Two studies compared modelling techniques in the development of their RATs (62,63). Wilson et al. reported only the logistic regression results but stated that Cox proportional hazards models had produced near identical results and were therefore not displayed (63). Heikes et al. observed equal accuracy for models based on logistic regression and a decision tree, opting to use the decision tree because of its 'greater ease of use' (62).

RATs tend to use data from questionnaires or routine non-invasive measures, this is good as it should ensure the RATs are considered acceptable by most individuals (56,59). The RATs which considered biochemical measures in addition to non-invasive measures showed they tend to slightly improve performance; while genetic profiling presently appears to be of little use (54,56).

Categorisation of continuous risk predictors was common with Collins et al. reporting 49% of the RATs categorised all continuous variables (57). However this practice is not advised, as it has been shown to lead to loss in the predictive power of the models and more worryingly severe bias when 'optimal' cut-points for the categories are derived using the data (64). On the other hand, Buijsse et al. highlight that the complexity of methods used are limited by the context in which the resulting RAT will be completed (56). The reason for the categorisation of continuous risk predictors in some cases is due to the fact that this allows the RAT to be completed in practice as a paper questionnaire. The majority of RATs did not consider nonlinear terms, despite the fact that it is advised when a nonlinear relationship between a risk factor and the outcome is observed, this may again be due to simple RATs being required (57,64). Though, with technology, for example smart phone apps, to calculate more complex RATs becoming more widely available, RATs may be allowed to develop further with the emphasis shifting from simplicity to accuracy (56).

The treatment of missing data in the datasets used to develop RATs is an important methodological issue, yet it was often not discussed in the RAT development papers (40,57). Of those that detailed the treatment of missing data the vast majority used a complete-case analysis, this is where any individuals with missing data are excluded from the dataset used to develop the RAT. Few RATs used multiple imputation which has been found to have more

valid results and better discrimination than the complete-case approach (65). Multiple imputation uses all the observed values on an individual to assign plausible values for the missing data, this process is repeated a number of times, leading to a number of datasets which are then used to make several risk models and an average is taken to determine the risk score. If a complete-case analysis is used its good practice to carry out a sensitivity analysis to assess the potential impact of missing data, although few did this.

Variables to include in the RATs were frequently chosen using an automated selection method, such as forward selection or backward elimination for regression models (57). Many of these automated approaches are criticised for being prone to over-fitting, where the performance seen for the data the model was developed from is over optimistic (66,67). Though there are more sophisticated stepwise techniques which have been developed to avoid these issues, such as Least Absolute Shrinkage and Selection Operator (LASSO) and Least Angle Regression (LAR) (68), these have yet to gain popularity with those developing RATs. Moreover it has been pointed out that any completely automated approach will encounter problems as they fail to take into account the context of the specific situation (69). Variables should be selected using statistical methods alongside expert clinical knowledge as well as previous evidence (66); as seen in the development of FINDRISC where physical activity and vegetable or fruit consumption which were not significant predictors were included to emphasise their importance in diabetes prevention (42). Finally, several RATs firstly carried out bivariate tests of association of candidate independent variables with the outcome to reduce the number of independent variables (57). This method is unwise as it may well remove variables that are good predictors of the outcome once another variable has been adjusted for (70,71).

Calibration and discrimination are two important measures of a RAT's utility (40). The vast majority of RATs reported discrimination, normally using AUROC, for at least one of the following: the data used to derive the model, an internal validation dataset or an external validation dataset (54,56-59). On the other hand, calibration was frequently not reported in development papers (54,56-59). The Hosmer-Lemeshow goodness-of-fit test was the most commonly used

measure of calibration (57). However Hosmer-Lemeshow has been shown to give misleading results in some situations, therefore it has been recommended that calibration plots should be preferred (72-74).

Common ways to internally validate were resampling methods or by cross-validation, where the dataset is split into a training set used for model development and a test set used to test the validity. However cross-validation has been criticised for being imprecise and thus it has been suggested that resampling methods such as bootstrapping or v-fold cross-validation are more robust ways to carry out internal validation (75).

External validation assesses the statistical performance of a RAT in a dataset independent to the dataset used to develop the RAT (76). Many RATs have not being externally validated and impact studies are particularly infrequent (54-59). External validations that have been carried out tend to show the performances of the RATs in external populations are poorer in terms of discrimination, sometimes to a concerning extent (77). Yet, the systematic review by Abbasi et al. (54) examined the external validity of 25 risk scores and found that the RATs' discrimination in their external dataset was comparable to the discrimination observed in the internal dataset; in some cases even being marginally better than that seen for the internal dataset, the authors suggest this could be due to higher heterogeneity in the external populations than the internal populations. The dataset used to externally validate the RATs in this review had a lower incidence of T2DM than most of the datasets used to develop these RATs, this resulted in poor calibration with significantly less observed cases than predicted by the prediction models. Recalibrating the prediction models, as done in this paper, by recalculating the intercept term in logistic regression models and the baseline survival function in survival models, deals with miscalibration due to differing incidence rates. This resulted in well calibrated RATs for this population, however poor calibration in external populations is not solely due to this issue and thus recalibration of intercept term or baseline survival function will not always result in a RAT which performs well (78).

Recalibration of the intercept term or baseline survival function is not the only way models can be required to be adapted when being externally validated. Noble et al. note that researchers externally validating RATs often had to remove variables from models as they were not in the external dataset. Where this happened it was common that an additional variable was added to account for the missing information; also common was altering the way in which categories for some variables were defined, for example ethnicity (58). Although these are still valid external validations as it assesses the performance of RATs in different settings, as Buijsse et al. highlight external performance is generally better in datasets which had the same variables definitions as the dataset on which the RAT was developed (56). Poor external performance is often seen when the demographic of the external population differs from that of the original population on which the RAT was developed, this being especially evident when ethnicities or countries of the populations differ (49). It has been suggested that re-estimating the existing model's coefficients in these situations could be a solution resulting in a model with satisfactory performance for this external population (79). However others state that ideally each population should have its own specific RATs developed, though this is expensive and unfeasible in many cases (80).

An impact study assesses the rate of uptake of a RAT and the subsequent interventions that are available in a real world setting, as well as whether having these available leads to improvement in outcomes. An example is the GOAL study which found that some individuals that are encouraged to change lifestyle will achieve it but most will have difficulty (81). The majority of RATs have not had studies into their impact in practice carried out; although current research is starting to focus on this area more, some of the RATs have been developed without thinking about their use in practice and thus are unlikely to be used (58,59). Interventions carried out in individuals identified as being at high risk of diabetes by RATs have achieved positive changes in risk factors, such as reduced weight; while these changes were statistically significant only some were clinically significant (82-84). One study's preliminary findings indicate that intervention in individuals detected by a diabetes risk score reduces incident diabetes in real world settings (85). Further research is needed to understand

the issues that inhibit and enable the use of self-assessment RATs in practice, as well as the cost and cost effectiveness of using RATs (58,59).

Since only three RATs with an outcome of prevalent NDH and undiagnosed T2DM have been included in these reviews, due to a lack of focus on RATs with this outcome, a systematic review focusing on RATs which include prevalent NDH in their outcome is included in this thesis. The findings that the basic decision tree may be unstable, means that extensions of the method which use resampled versions of the data are considered in the method comparisons chapter. Although this chapter considers the utility of methods and their variations in practice alongside statistically performance when comparing different methods, many RATs have been developed without thinking about their use in practice and thus are unlikely to be used (58,59). The methods comparison chapter will be an empirical comparison, using one dataset to develop RATs and another to assess the external validity, since RATs performance in external populations was often much worse than expected and external validation is the gold standard.

RATs developed in this thesis are built taking into consideration the way(s) they would be used in practice, as not to render them unusable in practice. Further to this RATs developed in this thesis will only use data available from questionnaires or routine non-invasive measures, as it should ensure the RATs are considered acceptable by most individuals (56,59). Finally RATs developed in this thesis which are intended to be implemented in practice will use expert clinical knowledge and previous evidence in addition to the statistical data as this is recommended (66).

1.5 Overview of thesis

The work reported by this thesis had two main aims. Firstly, to investigate methodological issues, informing the field of RATs as a whole. Secondly, to identify, develop and validate RATs for NDH and T2DM outcomes, thus helping to provide valid RATs which may be implemented in practice. The core chapters of this thesis which detailed the work carried out to meet these aims are summarised below in bullet point form.

- Chapter 3 presents a systematic review which identifies, summarises and assesses the methodology of RATs which detect those with NDH.
- Chapter 4 reports a comparison of existing methods for developing RATs, both in terms of implementation in practice and statistical performance. It also details a resampling study to assess the effects of differing the sample size of the development dataset on the performance of each of the methods.
- Chapter 5 describes the performance of utilising the chain event Graph (CEG) for the novel application of RATs.
- Chapter 6 assesses whether the LSA risk score should be replaced with a RAT which was developed with HbA1c as its outcome.
- Chapter 7 establishes risk groups for LPRS to enable consistent advice to be given across different general practices when utilising the RAT.
- Chapter 8 presents an external national validation of the LSA and LPRS using a nationally representative longitudinal dataset.

Chapter 2 details the datasets used for this work and statistical methods used to assess the performance of the RATs throughout this thesis. Finally Chapter 9 summarises the findings in light of the previous findings in the field.

Chapter 2: Datasets and statistical metrics

2.1 Chapter Outline

This chapter gives a summary of each of the three datasets which are analysed in this thesis. Additionally, the chapter details statistical metrics which are used to assess the performance of risk assessment tools (RATs) in this thesis.

2.2 Datasets analysed in this thesis

The three datasets used in the various analyses in this thesis are detailed below. Briefly the ADDITION-Leicester and Screening Those at Risk (STAR) datasets are cross-sectional datasets comprised of individuals from Leicestershire; while the English Longitudinal Study of Aging (ELSA) dataset is a longitudinal dataset which contains individuals from across England. Table 2.1 displays the analyses, by chapter, each of the datasets are utilised for.

Table 2.1 Datasets used for the analyses included in each chapter

	Dataset used for analyses		
Chapter	ADDITION-Leicester	STAR	ELSA
4 Methods for developing risk assessment tools using cross-sectional data and resampling study on the effects of the sample size of the development dataset on performance	√	√	
5 Chain event graphs for developing risk assessment tools using cross-sectional data	√	√	
6 Development of self-assessment risk assessment tools with HbA1c outcome and comparison of external validity for HbA1c outcome with LSA score	√	√	
7 Establishing risk groups for the Leicester Practice Risk Score	√	√	
8 Evaluating the ability of the Leicester Self-Assessment and Leicester Practice Risk Scores to identify individuals who go on to develop diabetes using longitudinal data			√

Datasets used to develop and validate RATs should firstly reflect well the population in which they are intended to be used, especially in terms of the variables known to be related to the outcome. Datasets should be cohorts rather than from case-control studies, with prospective cohorts being advocated as allowing optimal recording of risk factors and outcomes (86). Results are more likely to be reliable if large high quality dataset are used (40). Finally, datasets used for external validation should ideally be in a different location, providing geographic external validation (87).

2.2.1 ADDITION-Leicester

The ADDITION-Leicester dataset was yielded as part of the multi-centre ADDITION (Anglo-Danish-Dutch Study of Intensive Treatment in People with Screen Detected Diabetes in Primary Care) study (88). ADDITION-Leicester included population based blood screening tests, to identify individuals with undiagnosed Type 2 diabetes mellitus (T2DM) as well as those with non-diabetic hyperglycaemia (NDH). After blood screening tests individuals identified as having undiagnosed T2DM were invited into a randomised controlled trial of cardiovascular disease risk reduction; while individuals identified as having NDH, specifically impaired glucose regulation (IGR), were followed-up annually for five years to study the progression from NDH to T2DM. This thesis only uses data from the baseline screening phase of the ADDITION-Leicester study, which is detailed further below.

Twenty general practices from both urban and rural areas across Leicestershire were included in the study, resulting in a multi-ethnic dataset (88). All of these practices met a set inclusion criteria regarding the completeness of demographic data in their practice database. Individuals from these practices who did not have diagnosed diabetes and were aged 40-75 years old, or aged 25-75 years old for individuals from black and minority ethnic (BME) groups, were eligible for the study. Six of the practices invited all their eligible patients to participate in the screening phase of the study, while 14 practices invited random samples of their eligible individuals. Of the 30,950 individuals invited 6,749 (21.8%) attended the diabetes screening (89). Only 359 individuals were younger than 40 years old, due to the low proportion of individuals in this age group the analyses in this thesis exclude these individuals. The number of individuals included in each of the analyses varies, as exclusion criteria for the completeness of the data are used for some of the analyses.

The diabetes screening was carried out using a standard 75g Oral Glucose Tolerance Test (OGTT) (88). Individuals with either fasting plasma glucose (FPG) or 2-hour Glucose Tolerance Test (GTT) in the diabetic range (FPG ≥7.0 mmol/l or 2-hour GTT ≥11.1 mmol/l (90)), had a second glucose test to confirm diabetes at an additional visit within a week of the original visit. 206 (3.2%) of the individuals aged 40 years or older were found to have screen-detected

T2DM. The remaining individuals were classified as having either normal blood glucose, 80.5%, or NDH (IGR), 16.3%, according to the World Health Organisation (WHO) definitions (13,46).

The screening phase of the study collected the following data from participants: detailed family and medical histories, clinical measurements, anthropometric measurements, biomedical measurements and their answers to several validated questionnaires for a range of outcomes. The biomedical measures collected included glycated haemoglobin A1c (HbA1c), which has since been recommended for diagnosing T2DM (12). This enables ADDITION-Leicester to be used to assess whether the Leicester Self-Assessment (LSA) needs to be replaced with a new self-assessment RAT, developed with a binary HbA1c outcome, work which is carried out in Chapter 6. Furthermore, the validated questionnaires included two questionnaires for the purpose of predicting risk of diabetes, Finnish Diabetes Risk Score (FINDRISC) and Cambridge Risk Score (CRS) (42,48). Meaning values of potential risk factors were recorded for individuals in this dataset; which is helpful for developing potential risk scores as done with the dataset in Chapter 6.

ADDITION-Leicester is used in the analyses in Chapters 4, 5, 6 and 7 of this thesis. The numbers included in different analyses varies slightly, the characteristics of individuals aged 40 years and older in ADDITION-Leicester are displayed in Table 2.2.

.

Table 2.2 Characteristics of individuals aged 40 years and older who took part in diabetes screening for the ADDITION-Leicester study (n=6,390)

Age, years	57.3 (9.6)
Sex, Male (%)	47.7
Ethnicity, White European (%)	75.8
Body mass index (kg/m^2)	28.1 (5.0)
Waist (cm)	94.2 (13.1)
IGR or undiagnosed T2DM by OGTT (%)	19.6
HbA1c ≥6.0% (%)	23.2
Fasting plasma glucose (mmol/l)	5.2 (0.9)
HbA1c (%)	5.7 (0.6)
Systolic blood pressure (mmHg)	137.9 (19.4)
Diastolic blood pressure (mmHg)	85.7 (10.5)
Current smoker (%)	14.5
Used high blood pressure drugs: self-reported (%)	23.4
Medical record of antihypertensive medication use (%)	23.8
Cholesterol (mmol/l)	5.6 (1.1)
History of high blood pressure (%)	27.8
History of Angina (%)	4.8
1st Degree Relative with diabetes (%)	25.2

Mean (SD) displayed unless stated.

One advantage of this dataset is that it is a cohort of the general population with over a 20% participation rate meaning findings from using this dataset should be able to be generalised to the population. Furthermore, the dataset primarily consisted of 40- 75 year olds, the age group in which RATs are recommended for NDH and T2DM outcomes, meaning little data was lost when excluded individuals outside the age group of interest. The dataset is multi-ethnic, although the majority of individuals who participated who were not white European were of south Asian ethnicity. Additionally, both OGTT and HbA1c measurements were collected which is very rare amongst population screening studies. Many potential risk factors, including all well-known factors, were collected; however they did contain a reasonable amount of missing data. Finally, the size of the dataset was sufficient with a large ratio of NDH or undiagnosed T2DM to candidate predictors.

2.2.2 Screening those at risk (STAR)

The STAR study was a study of a targeted screening program which aimed to identify individuals with abnormal glucose tolerance, either NDH or undiagnosed T2DM (46). Individuals aged 40-75 years old, or aged 25-75 years old for individuals from BME groups, who did not have diagnosed diabetes and had at least one recognised risk factor for T2DM, were eligible to participate in the study. Eligible individuals from 17 general practices across Leicestershire were invited by letter to participate. Additionally, opportunistic recruitment of participants was also carried out at retail centres in Leicestershire during the "Be a star campaign" for health awareness. It is important to note that the general practices used to recruit individuals were different from those used for ADDITION-Leicester.

Participants required at least one of the following risk factors to be eligible for the study (46):

- Coronary heart disease
- Hypertension
- Dyslipidemia
- Cerebrovascular disease
- Peripheral vascular disease
- History of impaired glucose tolerance (IGT)
- Gestational diabetes
- First-degree relative with T2DM
- Body mass index (BMI) >25 kg/m^2
- Current or former smoker

Individuals with a terminal illness or who were housebound were excluded from the STAR study.

3,225 participants were screened for abnormal glucose using a 75g-OGTT. As in the ADDITION-Leicester study, the WHO 1998 criteria were used to diagnose abnormal glucose tolerance (T2DM or NDH (IGR)) (90). The diagnosis of T2DM also required a second confirmatory OGTT. During the screening visit trained research staff also collected biomedical and anthropometric data for each individual, and participants self-completed a general health questionnaire.

Finally the dataset includes patients' past medical and medication history, which was obtained by a qualified nurse.

STAR is used to externally validate the work carried out in Chapters 4, 5, 6 and 7 of this thesis. Although the number of individuals included in the different analyses varies as different exclusion criteria for the completeness of the data are used for some of the analyses. The participants under 40 years old (n=54) are excluded from every analysis in this thesis due to the small number of individuals of this age and since population level two-stage risk screening is only recommended for 40-75 year olds (28). Table 2.3 below gives a summary of the characteristics of the 3,173 individuals from STAR aged 40 years or older.

Table 2.3 Characteristics of individuals aged 40 years and older who took part in diabetes screening for the STAR study (n=3,173)

Age, years	56.6 (9.6)
Sex, Male (%)	46.4
Ethnicity, White European (%)	70.9
BMI (kg/m^2)	27.9 (5.1)
Waist (cm)	94.9 (13.0)
IGR or undiagnosed T2DM by OGTT (%)	21.4
HbA1c ≥6.0% (%)	29.2
Fasting plasma glucose (mmol/l)	5.3 (1.0)
HbA1c (%)	5.8 (0.7)
Systolic blood pressure (mmHg)	134.0 (20.7)
Diastolic blood pressure (mmHg)	80.4 (10.8)
Current smoker (%)	25.4
Antihypertensive medication use (%)	22.6
Cholesterol (mmol/l)	5.4 (1.0)
History of high blood pressure (%)	33.6
History of Angina (%)	5.5
1st Degree Relative with diabetes (%)	34.9

Mean (SD) displayed unless stated.

The key disadvantage of this dataset is that the study was carried out in the same location as the ADDITION-Leicester and thus it does not provide a geographical external validation only a temporal external validation for the analyses carried out in Chapters 4, 5, 6 and 7. Another disadvantage was that individuals required one known risk factor for T2DM to be included in the study, meaning this was not a population screening study. Although the risk profile was similar to the ADDITION-Leicester risk profile. Advantages of the dataset are it is a multi-ethnic cohort which primarily consisted of 40- 75 year olds. Furthermore, both OGTT and HbA1c measurements were collected along with

many potential risk factors, including all well-known factors. Finally, the size of the dataset was sufficient with a large ratio of NDH or undiagnosed T2DM to candidate predictors.

2.2.3 English Longitudinal Study of Aging (ELSA)

ELSA is an ongoing panel study of a nationally representative cohort of people aged 50 years and older living in England (91). ELSA collects data from participants every two years, each round of data collection is known as a wave. The first wave (conducted in 2002 and 2003) of the study recruited individuals living in private housing aged 50 years or older and their partners, irrespective of age, from households in which an individual participated in the Health Survey for England in 1998, 1999 or 2001 and agreed to follow-up (92). Refreshment samples of people aged between 50- 53 years were added to the study cohort in waves 3, 4 and 6 to ensure the cohort continued to be representative of individuals of this age. The study collects a broad range of data including demographic, economic, social, psychological, mental and physical factors along with various blood assays (92). The socio-demographics of the ELSA dataset have been found to be generally reflective of the English population (92).

Participants are asked to complete a questionnaire every wave, with nurse visits being conducted every other wave (every four years) to collect further information, such as blood test measurements (92). ELSA was purposely designed with the ability to study the prevalence of NDH and/or undiagnosed T2DM at waves 2, 4 and 6 as well as the incidence of self-reported diagnosed T2DM at every wave (93). A fasting blood glucose measurement and an HbA1c measurement is taken from willing individuals during nurse visits, every four years. Participants are asked at each wave whether they have ever been told they have diabetes by a doctor and if they are taking insulin or medication for diabetes. ELSA is a freely available dataset accessed through the United Kingdom (UK) Data Archive (94).

The analyses in Chapter 8 of this thesis use this dataset to assess the performance of the LSA and Leicester Practice Risk Score (LPRS), which were developed for cross-sectional outcomes, for both cross-sectional outcomes of NDH or undiagnosed T2DM at baseline and longitudinal outcomes of diabetes incidence within four and eight years. Wave 2 (conducted in 2004 and 2005) is taken as the baseline, as this is the first wave which included nurse visits and therefore blood measurements. 9,432 individuals participated in wave 2 of the

study (92). The analyses in Chapter 8 include differing numbers of individuals as the use of the risk score is assessed for many differently defined diabetes outcomes, as well as being assessed on its own and as part of a two-stage screening programme. To be included in any of the analyses individuals had to be aged 50-75 years old and free from diagnosed diabetes in wave 2, Table 2.4 displays the characteristics of the 6,778 individuals in this age range and free from diabetes diagnosis in wave 2.

Table 2.4 Characteristics of individuals aged 50-75 years and free from diabetes at Wave 2 of the ELSA study (n=6,778)

Age (years), Median (IQR)	61 (57-67)
Sex (Male)	44.5%
Ethnicity (White European)	98.0%
BMI (Kg/m^2)	27.8 (4.7)
Waist circumference (cm)	94.7 (12.8)
Family History of diabetes (at eight year follow-up)	14.2%
History of high blood pressure	38.5%
Antihypertensive medication use	12.5%
Glycated haemoglobin (%)	5.5 (0.5)
Fasting plasma glucose (mmol/l)	4.9 (0.7)
Systolic Blood Pressure (mmHg)	132.9 (17.7)
Diastolic Blood Pressure (mmHg)	76.3 (10.5)
Blood total cholesterol (mmol/l)	6.1 (1.1)
Current Smoker	14.6%

Mean (SD) displayed unless stated.

This study has the benefit of being designed to be nationally representative of older individuals in the UK, meaning it provides a dataset for geographical external validation of the LSA and LPRS. Additionally the study is a prospective cohort meaning the validity of the two RATs could be assessed for longitudinal T2DM outcomes, in addition to the cross-sectional NDH or undiagnosed T2DM outcomes. One issue with the dataset is that it does not contain any individuals in their forties, meaning the whole age range in which screening is recommended could not be assessed. The size of the dataset was good, although some risk factors of the LPRS and LSA had missing data. A particular weakness was family history of diabetes was not collected at baseline and had to be imputed from the final follow-up.

2.3 Statistical methods for assessing risk assessment tools

The core statistical methods used across this thesis are detailed in this section. Specifically, how to calculate each of the statistical metrics along with what exactly it is that they are used to assess in the context of this thesis and in which chapters of the thesis they are employed.

2.3.1 Statistical methods for assessing binary screening decisions

Sensitivity, specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are all statistical metrics which evaluate the performance of binary predictions for a specified binary outcome. Prevalence of the disease, proportion correctly classified and proportion classified with high risk are also statistics of interest when considering the performance of binary screening decisions, and thus will also be detailed. In the case of this thesis, these metrics are used to assess how well screening decisions given by a defined threshold of a risk score match the actual outcome of interest, the true glucose status of individuals.

Table 2.5 indicates the four possible situations which can arise from the use of binary screening decisions (95). The letters *a*, *b*, *c* and *d* are used to notate the number of individuals in a dataset which fall into each of the situations. Firstly, individuals who have the disease and are correctly screened as positive are known as true positives; the number of true positives is labelled *a*. However, individuals may also be screened incorrectly as positive when they do not have the disease, these are termed false positives; the number of false positives is notated by *b*. Another possibility is that individuals without the disease are correctly screened as negative, such individuals are counted as true negatives; the number of true negatives is labelled *c*. Finally, individuals with the disease can be incorrectly screened as negative, these are known as false negatives; the number of false negatives is notated by *d*.

Table 2.5 The four possible situations resulting from using a binary screening decision to screen for a binary outcome with commonly used letter-notation also indicated

		True disease status		
		Disease	No disease	Totals
Screening decision	Positive	**True positives** a	**False positives** b	a+b
	Negative	**False negatives** c	**True negatives** d	c+d
	Totals	a+c	b+d	a+b+c+d

Thresholds for deciding binary screening decisions should be chosen to sensibly balance the number of individuals in the four possible situations given the context. The statistical metrics detail to what extent that happens, giving the proportion of individuals from a combination of the situations being in one (or two) situation(s). Equations (2.1) to (2.7) detail how each of the statistical metrics are calculated using the notations introduced in Table 2.5. Sensitivity is the proportion of individuals with the disease who correctly receive a positive screening decision (96). Specificity is the proportion of individuals without the disease who are correctly given a negative screening decision (96).

$$\text{Sensitivity} = \frac{a}{a+c} \tag{2.1}$$

$$\text{Specificity} = \frac{d}{b+d} \tag{2.2}$$

As equation (2.3) details, PPV is the proportion of individuals given a positive screening decision who actually have the disease (97). Similarly NPV is the proportion of individuals given a negative screening decision who are actually free from the disease (97). PPV and NPV are of interest to patients who have received the result of a screening test, since they are the probabilities of having the disease given screening positive and not having the disease given screening negative respectively.

$$\text{PPV} = \frac{a}{a+b} \tag{2.3}$$

$$\text{NPV} = \frac{d}{c+d} \tag{2.4}$$

The prevalence of the disease is the proportion of individuals who have the disease out of all individuals screened. Changes to the prevalence directly result in changes to the predictive values with increasing prevalence increasing PPVs and decreasing NPVs are produced (95). While, in theory the sensitivity and specificity of a screening decision does not vary with changing levels of prevalence.

$$\text{Prevalence } = \frac{a+c}{a+b+c+d} \tag{2.5}$$

As displayed by (2.6), the proportion of individuals correctly classified is the sum of the true positives and true negatives divided by the total number of individuals screened. The proportion classified as high risk is the total number of individuals with a positive screening decision over the total number of individuals screened.

$$\text{Proportion correctly classified} = \frac{a+d}{a+b+c+d} \tag{2.6}$$

$$\text{Proportion classified as high risk} = \frac{a+b}{a+b+c+d} \tag{2.7}$$

All the statistical metrics detailed here are commonly reported as both proportions, as shown in equations (2.1) to (2.7), and percentages yielded by multiplying these proportions by 100. They are reported throughout the thesis, along with 95% confidence intervals (CIs).

2.3.2 Area under the receiver operator characteristic curve

The receiver operator characteristic (ROC) curve is a plot of the true positive rates against the false positive rates yielded by taking each observed risk score as the cut-point which determines the screening decisions (98). As the example in Figure 2.1 indicates, the true positive rate is simply the sensitivity of a particular cut-point; while the false positive rate is one minus the specificity. The area under the ROC is the area between the curve and the lines $x = 1$ and $y = 0$.

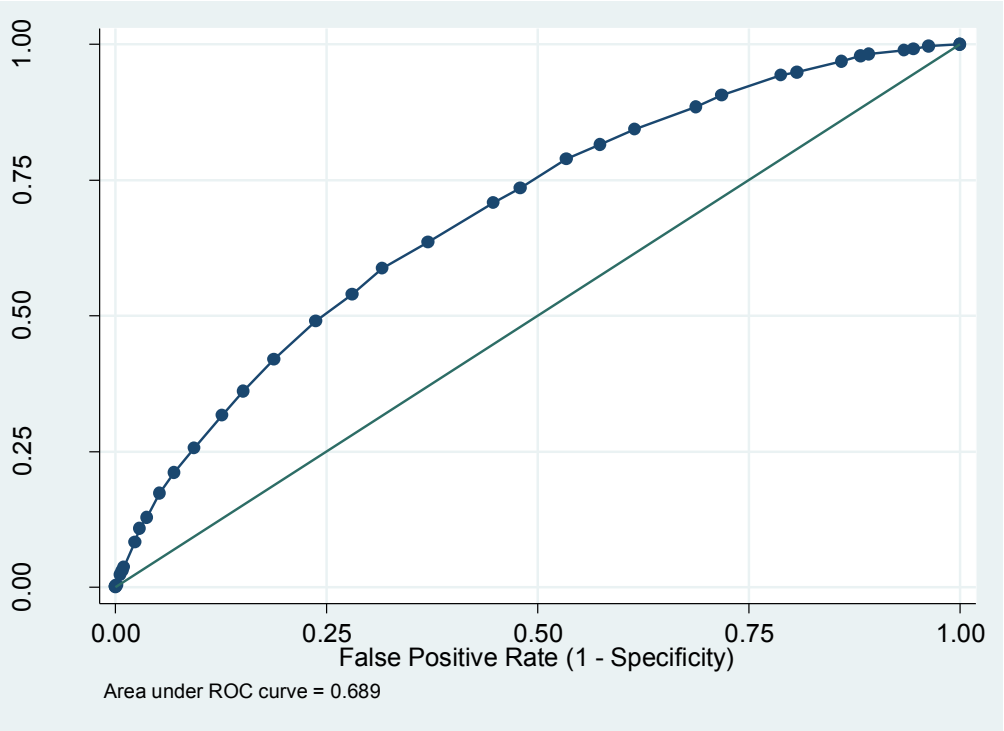


Figure 2.1 Example of a ROC curve

The area under the receiver operator characteristic curve (AUROC) is the probability that a randomly selected individual with the disease will have a higher risk score than a randomly selected individual without the disease (98). The AUC is a commonly used statistical metric for assessing the discrimination of a risk score (54-59,99).

As in Figure 2.1 it is common to display a line from the bottom left-hand corner to the top right-hand corner as a reference when plotting the ROC curve (98). This is known as the line of no discrimination, since it shows where the ROC curve would lie if its discrimination is no better than chance. The area under the line of no discrimination is 0.5, indicating that only half the time a randomly

selected individual with the disease will have a higher risk score than a randomly selected individual without the disease. On the other hand, a risk score with perfect discrimination would go from the bottom left-hand corner to the top left-hand corner and then across to the top right-hand corner. The area under a perfect ROC curve would be 1, since every individual with the disease has a higher risk score than every individual without the disease.

2.3.3 Brier score

Brier scores measure the accuracy of the predicted probabilities of the outcome. The Brier score is often thought of as a measure of calibration and was chosen to assess calibration in this thesis. However the Brier score measures overall fit of a models, of which calibration is a component rather than being a measure of calibration alone (100,101). The Brier score was chosen since it is more consistent than tests of perfect calibration, such as the Hosmer-Lemeshow test, which has historically been used (73). Plots of observed against predicted risk by decile are also used in the thesis to assess calibration.

The Brier score is the average of the sum of squared errors in predicted probabilities of each classification and whether that classification was observed for each individual. Since this thesis only predicts binary classifications the Brier score can be simplified to equation (2.8); where f_{ij} is the predicted probability of outcome j for individual i, and E_{ij} is an indicator of whether event j was the true outcome for individual i (taking 0 when j is not the true outcome and 1 when j is the true outcome). The Brier score is a value between 0 and 1, the closer the value is to 0 the better the calibration of the model.

$$\text{Brier Score} = \frac{1}{N}\sum_{j=1}^{2}\sum_{i=1}^{n}\left(f_{ij} - E_{ij}\right)^2 \tag{2.8}$$

The Brier score is affected by the prevalence of the outcome, with a decrease in the prevalence leading to a decrease in the outcome index variance, which is the Brier score yielded from assigning each individual the prevalence as their prediction of the outcome (102). The outcome index variance is used as a reference for the Brier scores displayed in this thesis, since it gives the Brier score yielded from the non-informative model.

Some risk scores, such as those intended to be calculated by hand, may not be communicated as a probability due to making their calculation easy for a lay person to carry out, in their cases if possible the Brier score is calculated using the associated probability from the underlying model for each risk score.

2.3.4 Net reclassification improvement

Net reclassification improvement (NRI) measures the extent to which a new model correctly reclassifies individuals into risk categories for a binary outcome compared to the risk categories they were allocated using an existing model (103). As (2.9) shows NRI is the sum of the NRI for the events, NRI_E, and the NRI for the non-events, NRI_{NE}. In the context of this thesis events are individuals with the outcome of interest, while non-events are individuals without the outcome of interest. As expressed in (2.10), NRI_E is the probability of events correctly moving up the risk categories minus the probability of events incorrectly moving down the risk categories when reclassified using the new model compared to the existing model. (2.11) indicates NRI_{NE} is the probability of non-events correctly moving down the risk categories minus the probability of non-events incorrectly moving up the risk categories.

$$\mathrm{NRI} = NRI_E + NRI_{NE} \tag{2.9}$$

$$NRI_E = P(up|event) - P(down|event) \tag{2.10}$$

$$NRI_{NE} = P(down|non\ event) - P(up|non\ event) \tag{2.11}$$

Chapter 3: Systematic Review of risk assessment tools for detecting those with non-diabetic hyperglycaemia

3.1 Chapter Outline

This chapter identifies, summarises and assesses the methodology of risk assessment tools (RATs) which detect those with non-diabetic hyperglycaemia (NDH).

The work in this chapter has been:

- Presented as a poster:
  Barber SR, Davies MJ, Khunti K, Gray LJ. 'Risk assessment tools for detecting those at high risk of type 2 diabetes: a systematic review'. At: Diabetes UK Professional Conference 2014. Liverpool, UK. 5-7th March 2014.
- Published:
  Barber SR, Davies MJ, Khunti K, Gray LJ. Risk assessment tools for detecting those with pre-diabetes: A systematic review. Diabetes Research and Clinical Practice 2014; 105(1):1-13.

3.2 Introduction

As stated in section 1.2, there are interventions that are effective in delaying and even preventing individuals with NDH progressing to Type 2 diabetes mellitus (T2DM). It has been shown the cost per case of detecting individuals with NDH is reduced by using RATs as part of the screening programme (104). However, as highlighted by section 1.4.3, a systematic review of RATs for detecting those with NDH has not been carried out, thus this chapter reports such a systematic review. In identifying, summarising and evaluating the RATs developed which detect those with NDH; the systematic review detailed in this chapter aids those that desire to use such a RAT in their selection or development of an appropriate RAT.

3.2.1 Search Strategy

The search strategy was devised to identify all articles which developed new RATs for the outcome of NDH with or without undiagnosed T2DM. The search strategy contains two sets of terms; one set of terms for the outcomes of interest, which includes terms for T2DM as well as NDH, and another set of terms for RATs. Articles had to contain a term from both sets to be considered for inclusion. The terms included in both sets were chosen after examining the terms used in the previous systematic reviews in the area as well as adding additional terms from the literature used for NDH. Two electronic sources, Medline and Embase were both searched from inception until the 7th January 2013 using a specific search strategy, which is given in Appendix A. In addition, reference lists of relevant articles were also manually searched. In-process or un-indexed work which was available on Medline was included in this review; however conference abstracts were not considered as the full article was required.

3.2.2 Inclusion/Exclusion Criteria

The outcome of a RAT for inclusion had to be one of the following:

1. NDH defined using oral glucose tolerance test (OGTT) i.e. impaired glucose regulation (IGR), impaired fasting glucose (IFG) and/or impaired glucose tolerance (IGT). (90)

2. NDH defined using glycated haemoglobin A1c (HbA1c) using any recommended definition. (14,28,29)
3. 1 and 2 from above, i.e. NDH by both OGTT and HbA1c.
4. 1 or/and 2 and current undiagnosed T2DM (by HbA1c or using OGTT).

The RATs had to contain two or more risk factors. Articles only assessing associations were excluded. RATs that included genetic factors were excluded as they are not routinely available in clinical practice and are expensive and time-consuming to collect so do not offer the same benefits over diagnostic tests that other non-genetic RATs do (54,56). The RATs had to be developed either on a population-based sample or volunteers/opportunist sample, i.e. not a pre-screened sample, for example individuals at an obesity clinic. In order for the methodology to be assessed the article had to detail the development of the RAT. Finally this review was restricted to articles published in English.

3.2.3 Article Selection

I examined the titles and keywords section of the articles identified by the search and excluded those which were not on the topic of interest. I then read the abstracts of the remaining articles and again excluded articles based on the inclusion criteria; any articles which appeared potentially relevant were kept for further examination. Papers which I did not have full access to were requested at this stage. The remaining articles were then examined fully (the whole text) by myself and a second reviewer, Dr Laura Gray, with any discrepancies being resolved through discussion.

3.2.4 Data Extraction

Data was extracted using a standardised data extraction form, given in Appendix A, to ensure consistent information was collected for each RAT; it included a series of questions examining the methodology. Information on the treatment of missing data, the number of outcome events and populations the RATs were developed and validated (if in the same paper) on were collected. Data extraction for internal and external validations included the area under the receiver operator curve (AUROC), sensitivity and specificity to summarise the discriminative ability of the RATs at the optimal cut point and any calibration statistics reported. Finally the risk factors considered and those included in the

RATs, as well as the mechanism by which the RAT could be used in practice, were recorded.

In addition to the data extraction from the paper, google scholar was used to determine the number of times each paper had been cited in other articles, this search was carried out on 27th June 2014. These articles were then examined to see whether impact studies, studies which assess the effect implying a RAT in practice, or external validation studies have been carried out.

3.2.5 Analysis

Simple statistics, such as medians and ranges, were calculated for size of datasets, AUROCs, number of risk factors considered, number of risk factors included and the number of events per candidate predictor variable (EPV). A candidate predictor variable is a variable that was considered for inclusion into the RAT at any stage.

3.3 Results

Figure 3.1 shows the number of articles considered at each stage of the systematic review. After examining titles and then abstracts, 256 full papers were reviewed, from which 12 articles met the inclusion criteria. These 12 papers, by 11 different authors, contained 18 RATs from which data were extracted.

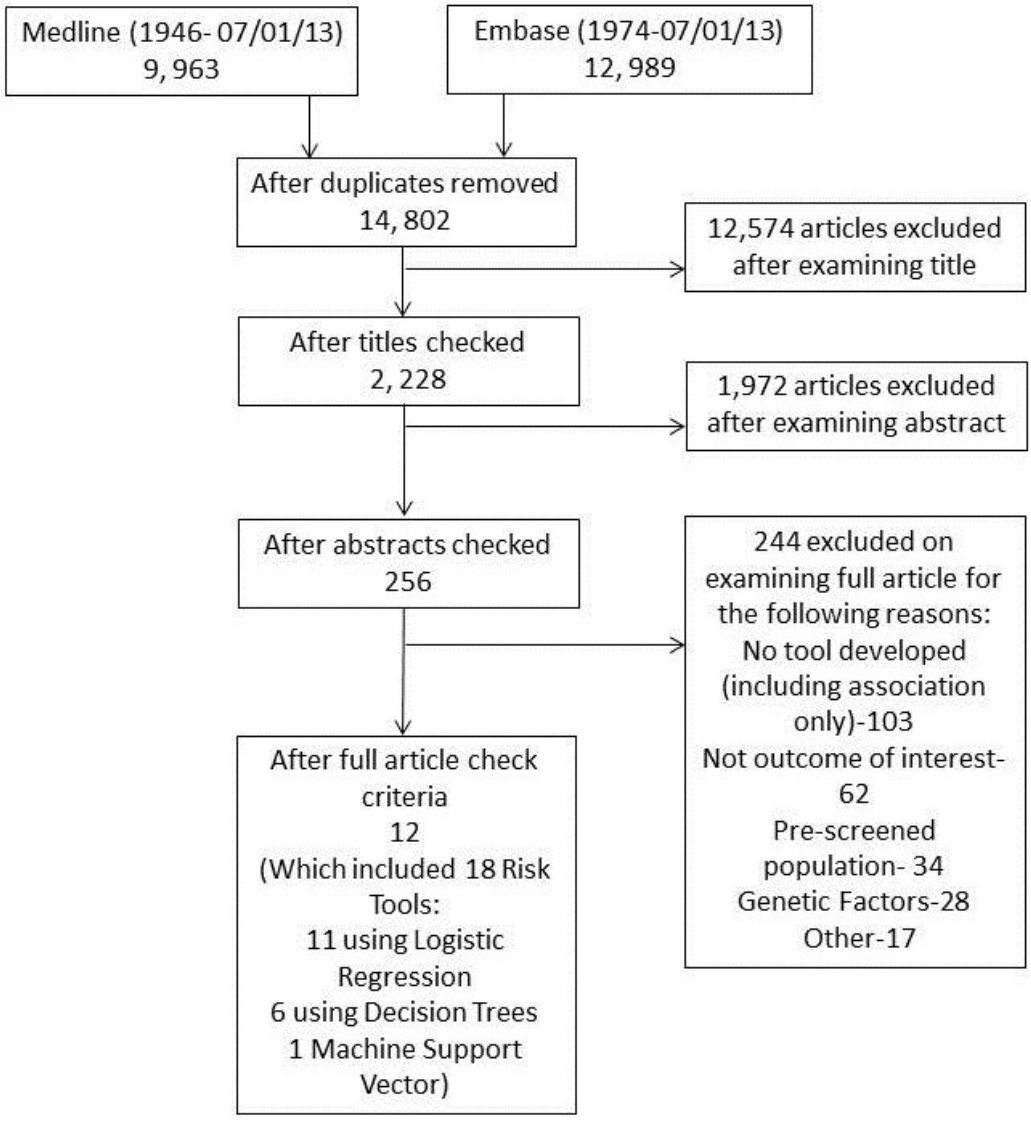


Figure 3.1 Diagram summarising paper selection

Table 3.1 shows that six RATs were developed using data from the United States of America (USA), four with data from the Middle East, two using data from each of the United Kingdom (UK), Canada and Germany and one using data from China. The majority of papers containing these RATs have been published in the last five years, with Barriga et al.'s 1996 article being the only paper published before 2000. Generally the RATs' outcomes were defined using OGTT measurements, either fasting plasma glucose (FPG), 2-h post challenge blood glucose or both. However the four RATs published in Handlos et al. use HbA1c levels to define their outcomes (105) and the computer-based RAT by Gray et al. uses HbA1c measurements along with FPG and 2-h post challenge blood glucose measurements to define its outcome (45).

3.3.1 Methods used to develop risk assessment tools

Of the 18 RATs identified 11 (61%) used logistic regression to develop the RAT, six (33%) used decision trees and one (6%) used a support vector machine (SVM). Logistic regression and decision tree based RATs allocated scores as described in section 1.4. SVM uses multidimensional hyperplanes to split continuous variables in relation to other variables (continuous or categorical) in order to divide the outcome variable into groups of mainly events and other groups of mainly non-events (106). This method is explained in more detail in section 4.7 in the following chapter.

3.3.2 Methodological quality

Ten (55.6%) of the RATs were developed using population-based data, with the remainder developed using data collected either through advertising or opportunistic sampling.

The risk factors considered for inclusion were listed in all but one article (107). The number of risk factors considered ranged from six to 26 with a median of 16. The number of outcome events in the datasets used to develop the RATs ranged from 244 to 2156 with a median of 644 (Table 3.1), giving events per variable (EPV) ranging from 15.3 to 128.4 with a median of 56.8. The final RATs contained between two and 19 risk factors with a median of six being included. The RATs developed using logistic regression had a median of 12 variables (range 4-19), while the decision tree scores had a median of three variables (range 2-5) and the SVM score included 11 variables. Figure 3.2 shows the

number of times variables were included in RATs compared to the number of times they were considered for inclusion using either the logistic regression or decision tree method. It can be seen in Figure 3.2 for example, that sex was included in nine of the 11 risk scores developed using logistic regression, however it was not included in any of the decision tree models. Age was the most frequently included variable for both methods. The SVM included ten of the 14 variables it considered; smoking, alcohol use, education and household income being the variables considered but not included. The variables included in each RAT are displayed in Appendix A.

Table 3.1 Summary of risk scores included in this systematic review

First author, year and citation number	Country /countries	Name of study used to develop score	Name of risk assessment tool	Difference to other tools by same authors	Intended /suggested use	Sampling frame	Method used to develop score	Sample size (number of events)	EPV	Outcome	Treatment of missing data
Barriga, 1996 (108)	USA	SLVDS (San Luis Valley Diabetes Study)	N/A	Simultaneous approach: all variables (telephone interview, clinical and blood sample) considered for inclusion in RAT.	By healthcare professional	20–74 year olds (residents of Alamosa and Conejos, English or Spanish speakers)	Decision Tree	1351 (244)	15.25	2-h post challenge glucose ≥140 mg/dL	Not mentioned
Barriga, 1996 (108)	USA	SLVDS	N/A	Stage 1 of phased approach: RAT only considers variables that can be collect over the telephone to detect individuals likely to have outcome.	In stages (starting with over telephone)	20–74 year olds (residents of Alamosa and Conejos, English or Spanish speakers)	Decision Tree	1351 (244)	15.25	2-h post challenge glucose ≥140 mg/dL	Not mentioned
DuBose, 2012 (109)	USA	NHAMES (National Health and Nutrition Examination Survey)	TAG-IT-A	N/A	Community-based screening	12–18 year olds	Logistic Regression	3050 (482)	80.3	Fasting blood glucose at or greater than 100 mg/dL, defined by ADA as impaired	Complete case

First author, year and citation number	Country /countries	Name of study used to develop score	Name of risk assessment tool	Difference to other tools by same authors	Intended /suggested use	Sampling frame	Method used to develop score	Sample size (number of events)	EPV	Outcome	Treatment of missing data
Gray, 2010 (51)	UK	ADDITION-Leicester	Leicester risk assessment score	To be used by layperson	Self-assessment in practice or community setting	40–75 year olds from a multi-ethnic UK screening study	Logistic Regression	6186 (1249)	56.8	Fasting blood glucose ≥6.1 mmol/l and/or 2 h blood glucose ≥7.8 mmol/l	available data analysis (with sensitivity analysis)
Gray, 2012 (45)	UK	ADDITION-Leicester	NS	To be used by professional with data available in primary care	In primary care using electronic database	40–75 year olds from a multi-ethnic UK screening study	Logistic Regression	6390 (1412)	128.4	Fasting blood glucose ≥6.1 mmol/l and/or 2 h blood glucose ≥7.8 mmol/l and/or HbA1c ≥ 6.5% (46 mmol/mol)	Not mentioned
Handlos, 2013 (105)	Algeria, Saudi Arabia and UAE	NS	NS	Regional score for Middle East & North Africa	Various settings	30–75 year olds offered screening in a central location of 6 cities	Logistic Regression	6588 (1173)	55.9	HbA1c ≥ 6.0% (42 mmol/mol)	'No' imputed for missing values

First author, year and citation number	Country /countries	Name of study used to develop score	Name of risk assessment tool	Difference to other tools by same authors	Intended /suggested use	Sampling frame	Method used to develop score	Sample size (number of events)	EPV	Outcome	Treatment of missing data
Handlos, 2013 (105)	Algeria	NS	NS	National score for Algeria	Various settings	30–75 year olds offered screening in a central location of 2 cities	Logistic Regression	2155 (386)	18.4	HbA1c ≥ 6.0% (42 mmol/mol)	'No' imputed for missing values
Handlos, 2013 (105)	Saudi Arabia	NS	NS	National score for Saudi Arabia	Various settings	30–75 year olds offered screening in a central location of 2 cities	Logistic Regression	2446 (377)	18.0	HbA1c ≥ 6.0% (42 mmol/mol)	'No' imputed for missing values
Handlos, 2013 (105)	UAE	NS	NS	National score for UAE	Various settings	30–75 year olds offered screening in a central location of 2 cities	Logistic Regression	1987 (407)	19.4	HbA1c ≥ 6.0% (42 mmol/mol)	'No' imputed for missing values
Heikes, 2008 (62)	USA	NHANES III	Diabetes Risk Calculator	N/A	Self-assessment using paper-based version and in clinical practice using electronic version	≥20 years old	Decision Tree	7092 (2156)	119.8	FPG ≥ 100 mg/dL or 2-h OGTT ≥ 140 mg/dL	Used FPG alone when 2-h OGTT missing

First author, year and citation number	Country /countries	Name of study used to develop score	Name of risk assessment tool	Difference to other tools by same authors	Intended /suggested use	Sampling frame	Method used to develop score	Sample size (number of events)	EPV	Outcome	Treatment of missing data
Hische, 2010 (110)	Germany	Mesy-Bepo (Metabolic Syndrome Berlin Potsdam Study)	N/A	Just clinical explanatory variables	Clinical practice	Individuals >18 years old from the cities of Berlin and Potsdam and the surrounding area	Decision Tree	1737 (601)	46.2	2-h post challenge glucose ≥140 mg/dL	Not mentioned
Hische, 2010 (110)	Germany	Mesy-Bepo	N/A	Clinical & laboratory explanatory variables	Clinical practice	Individuals >18 years old from the cities of Berlin and Potsdam and the surrounding area	Decision Tree	1737 (601)	23.1	2-h post challenge glucose ≥140 mg/dL	Not mentioned
Koopman, 2008 (111)	USA	NHANES	TAG-IT	N/A	Clinical and population settings, or to identify potential participants for research	20–64 year olds	Logistic Regression	4045 (1117)	111.7	Fasting blood glucose at or greater than 100 mg/dL	Not mentioned

First author, year and citation number	Country /countries	Name of study used to develop score	Name of risk assessment tool	Difference to other tools by same authors	Intended /suggested use	Sampling frame	Method used to develop score	Sample size (number of events)	EPV	Outcome	Treatment of missing data
Nelson, 2003 (112)	USA	NHANES III	NS	N/A	By clinicians	40–74 year olds	Logistic Regression	2746 (686)	76.2	2-h postchallenge glucose ≥140 mg/dL	Used BMI if triglyceride missing
Robinson, 2011 (113)	Canada	NS	eCANRISK	Electronic	Self-assessment & clinical in Canada (by electronic device)	Adults from seven provinces (most 40–74 years old, multi-ethnic group)	Logistic Regression	4091 (1273) ['Test' part of 6223]	106.1	Fasting blood glucose ≥6.1 mmol/l and/or 2 h Blood glucose ≥7.8 mmol/l	Various techniques including mean imputation and 'no' imputation
Robinson, 2011 (113)	Canada	NS	pCANRISK	Paper-based	Self-assessment & clinical in Canada	Adults from seven provinces (most 40–74 years old, multi-ethnic group)	Logistic Regression	4091 (1273) ['Test' part of 6223]	106.1	Fasting blood glucose ≥6.1 mmol/l and/or 2 h blood glucose ≥7.8 mmol/l	Various techniques including mean imputation and 'no' imputation
Yu, 2010 (106)	USA	NHANES (1999–2004)	Diabetes Classifier (II)	N/A	Self-assessed or professional	≥20 years old	Support Vector Machine	Training part of 4915 (1709)	122.1	Fasting plasma glucose level ≥100 mg/dL	Not mentioned

First author, year and citation number	Country /countries	Name of study used to develop score	Name of risk assessment tool	Difference to other tools by same authors	Intended /suggested use	Sampling frame	Method used to develop score	Sample size (number of events)	EPV	Outcome	Treatment of missing data
Xin, 2010 (107)	China	Beijing Community Pre-Diabetes Study	N/A	N/A	By practitioners at general practices (especially in rural China)	≥35 years old	Decision Tree	893 (393)	N/A	FPG ≥ 5.6 mmol/l and 2-h post challenge glucose ≥140 mg/dL	Not mentioned

N/A – not applicable, NS – not stated.

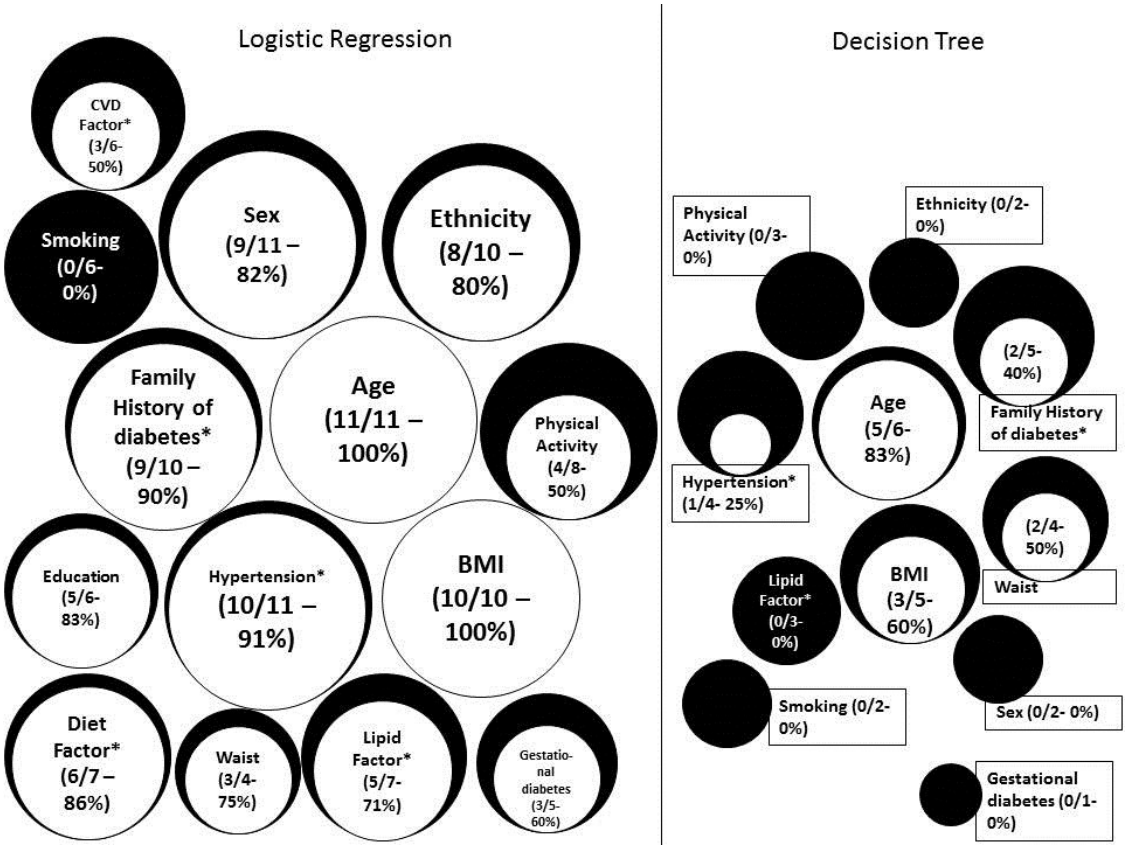


Figure 3.2 Frequency of variables included in tools (white) compared to number of times variable considered for inclusion in tools (black) split by the method used to develop the tool (logistic regression or decision tree)

Variables considered for inclusion in less than five RATs were not included in this figure. [*where several similar variables were considered for the same RAT (e.g. several different family history of diabetes variables), the white circle is frequency of at least one of the variables being included]

Of the RATs developed using logistic regression, six used a stepwise procedure to select variables for inclusion. One included all factors that were significant in univariate analysis (109). The two scores developed by Robinson et al. (113) used the full model of all factors available and the two other RATs used non-conventional methods to choose factors to include (111,112).

The RATs developed using decision trees used CART algorithms (60) or Quilan's standard decision tree algorithm (114) to build their tree. Two of these were chosen after comparison with a logistic regression model which was also developed in each article (62,107); they both stated this was due to its simplicity and similar accuracies as the logistic regression model. Although, simple

logistic regression based RATs have been developed in other papers (51,113), which allowed them to be paper-based, only requiring simple addition. A comparison of a simplified logistic regression based RAT with decision tree RAT would be useful to see if the logistic regression method is still comparable with the decision tree method after the scoring system for the RAT developed using logistic regression has been simplified.

The SVM method was also compared to logistic regression in one paper (106). Yu et al. used a SVM RAT as they thought the lack of parametric assumptions which exist in logistic regression would mean the method may perform well in detecting NDH and undiagnosed T2DM, which has complex relationships with its risk factors. However the SVM RAT and logistic regression RAT had a similar performance, the discrimination of the SVM model and logistic regression model were not significantly different (106).

Ten (91%) of the RATs developed using logistic regression categorised all continuous risk factors (51,105,109,111-113), with only one (45) opting to preserve continuous risk factors. Of the ten scores which categorised continuous variables, two were developed to be paper-based, thus difficult calculations involving continuous variables needed to be avoided. Six others stated they could be used in a variety of settings, including community use or population settings, which could also require them to be paper-based.

The treatment of missing data was not mentioned for eight (44%) of the RATs developed. Table 3.1 shows, the RATs which commented on missing data used an array of methods; including complete case analysis, ‘no’ imputation, mean imputation and a combination of techniques for different variables.

3.3.3 Validation

Table 3.2 and Table 3.3 display all RATs which included a validation in the same paper in which they were developed. Seven (39% of) RATs were validated using an external dataset (45,51,62,109-111). Fourteen (82%) of the RATs were validated internally using resampling techniques such as bootstrapping and cross-validation (45,62,105-108,110,112,113). Four of the RATs also randomly split their dataset into a ‘training’ set, which was used to create the RAT, and a ‘test’ set, which was used to validate the RAT

(62,106,113). Finally Xin et al.'s RAT was validated using a partially independent dataset (107); all cases were used in both the 'training' and 'test' cohorts however the non-cases were randomly split into these two cohorts. The reason given for this was a lack of cases in the dataset.

Table 3.2 Main features of the logistic regression risk scores

Risk assessment tool	Internal validation	Internal AUROC (95% CI)	Internal PPV of decision (%)	Internal NPV of decision (%)	Internal sensitivity of decision (%)	Internal specificity of decision (%)	Internal % needing further testing of decision	Calibration	External AUROC (95% CI)
DuBose (109)	N/A	NS	N/A	N/A	N/A	N/A	N/A	NS	0.61
Gray, 2010 (51)	Apparent	0.69 (0.68–0.71)	27.7 (26.2–29.3)	88.8 (87.7–89.9)	72.1 (69.6–74.6)	54.1 (52.7–55.5)	NS	NS	0.72 (0.69–0.74)[a]
Gray, 2012 (45)	Apparent	0.701 (0.684–0.717)	N/A	N/A	N/A	N/A	N/A	Hosmer–Lemeshow (***p*** = 0.97)	Ranged from 0.622 to 0.6851 for different outcomes[b]
Handlos (Middle East & North Africa score) (105)	Bootstrap	0.70 (for data from each of Algeria, Saudi Arabia & UAE)	NS	NS	Varies from 74 to 76 depending on sample	Varies from 50 to 54 depending on sample	Varies from 50 to 55 depending on sample	NS	N/A
Handlos (Algeria score) (105)	Bootstrap	0.70 (0.68–0.73)	NS	NS	74 (70–78)	57 (54–59)	49	NS	N/A
Handlos (Saudi Arabia score) (105)	Bootstrap	0.70 (0.67–0.72)	NS	NS	74 (70–78)	55 (53–57)	50	NS	N/A
Handlos (UAE score) (105)	Bootstrap	0.70 (0.67–0.72)	NS	NS	78 (73–82)	52 (50–55)	54	NS	N/A
Koopman (111)	N/A	0.74 (ranged from 0.717 to 0.753 when	N/A	N/A	N/A	N/A	N/A	NS	0.744

Risk assessment tool	Internal validation	Internal AUROC (95% CI)	Internal PPV of decision (%)	Internal NPV of decision (%)	Internal sensitivity of decision (%)	Internal specificity of decision (%)	Internal % needing further testing of decision	Calibration	External AUROC (95% CI)
		split by race groups)							
Nelson (112)	Bootstrap	0.74 (0.72–0.76)	N/A	N/A	N/A	N/A	N/A	NS	N/A[c]
Robinson (electronic score) (113)	'Test' part of split sample used to validate & Bootstrap	0.75 (0.73–0.78)	N/A	N/A	N/A	N/A	N/A	Chi-squared (**p** < 0.001), Brier Score (***p*** = 0.002), Hosmer–Lemeshow	N/A
Robinson (paper-based) (113)	'Test' part of split sample used to validate & Bootstrap	0.75 (0.73–0.78)	N/A	N/A	N/A	N/A	N/A	Chi-squared (**p** < 0.001), Brier Score (*p* = 0.002), Hosmer–Lemeshow	N/A

N/A – not applicable, NS – not stated.

[a] Gray et al. (51) was also externally validated in a youth South Asian population giving an AUROC of 0.71 (when using OGTT to define the outcome of NDH or T2DM) and an AUROC of 0.67 (when using HbA1c to define the outcome of NDH or T2DM) (115). In addition this RAT was externally validated in a Japanese cohort for the outcome of T2DM (either by FPG or HbA1c) where an AUROC of 0.804 was observed (116).

[b] Gray et al. (45) was also externally validated in a youth South Asian population giving an AUROC of 0.72 (when using OGTT to define the outcome of NDH or T2DM) and an AUROC of 0.68 (when using HbA1c to define the outcome of NDH or T2DM) (115). In addition this RAT was externally validated in a Japanese cohort for the outcome of T2DM (either by FPG or HbA1c) where an AUROC of 0.814 was observed (116).

[c] Nelson et al. (112) was externally validated by Piette et al. Honduran population for the outcome of FPG defined T2DM and had an AUROC 0.887 (117)

Table 3.3 Main features of the decision tree and SVM risk scores

Risk assessment tool	Internal validation	Internal AUROC	Internal PPV of decision (%)	Internal NPV of decision (%)	Internal sensitivity of decision (%)	Internal specificity of decision (%)	Internal % needing further testing of decision	Calibration	External AUROC (in same paper)
Barriga (Simultaneous approach) (108)	Tenfold cross-validation used to choose tree	0.73[a]	31	97	91	55	53	NS	N/A
Barriga (Stage 1 of serial approach) (108)	Tenfold cross-validation used to choose tree	0.665[a]	26	96	92	41	65	NS	N/A
Heikes (62)	'Test' part of Split sample used to validate & cross-validation	0.75	49	85	75	65	NS	NS	0.6991[b]
Hische (Clinical variables only) (110)	Tenfold cross-validation	0.668[a]	48.0	84.4	89.3	37.4	73.1	NS	0.6129[a]
Hische (Clinical & laboratory variables) (110)	Tenfold cross-validation	0.722[a]	56.2	89.1	89.7	54.6	67.3	NS	0.614[a]
Yu (106)	'Test' part of Split sample used to validate & tenfold cross-validation	0.739	67.3	80.9	70.9	65.9	NS	NS	N/A
Xin (107)	Partial split sample used to validate[c] & "Leave-one-out" method	0.689	N/A	N/A	N/A	N/A	N/A	NS	N/A

N/A – not applicable, NS – not stated.

[a] Calculated from the sensitivity and specific values stated

[b] Heikes et al.'s RAT was also externally validated in a different paper by a third party (AUROC: 0.67 for prediabetes, 0.70 for T2DM) (118). It was also externally validated for the outcome of T2DM (defined using either FPG or non-FPG levels) by Lee et al. in Korean datasets having AUROCs of 0.604 (KHANES 2001 and 2005) and 0.618 (KHANES 2007-2008) (119).

[c] All cases were used in both the 'training' and 'test' cohorts however the non-cases were split into these two cohorts.

3.3.4 Discrimination and calibration

Fourteen (78%) of the 18 RATs had an AUROC reported in their paper (45,51,62,105-107,109,111-113), the four RATs which did not were decision trees and they reported several values of specificity and sensitivity meaning an AUROC could be calculated from this information (108,110). The internal AUROCs of the RATs developed had a median of 0.7 (ranging from 0.66 to 0.75) (Table 3.2 and Table 3.3). Only three RATs had calibration measures reported, stating the Hosmer-Lemeshow goodness-of-fit test value (two of which also reporting the Brier Score and Chi-squared test value) (45,113). Gray et al. (45) is the only RAT to have a calibration plot published, with the observed number of cases being plotted against estimated number of cases for each decile of the dataset's predicted probability, giving a good visual demonstration that the RAT is well calibrated. However, Robinson et el.'s paper (113) presents a figure of the observed probabilities for each of the deciles dataset's risk scores, however it does not compare this to an estimated probability for each group.

3.3.5 Usability, impact studies and further external validation

All RATs either stated an intended or suggested use (Table 3.1). However, several of these were merely proposed ways in which it could be used rather than an exact situation and it was clear they did not take some of the development decisions with a particular use in mind. For example, Handos et al. (105) suggested that the RATs they developed were easy to implement in various situations, however did not state a specific setting which they intended their RAT to be used in. Some were, however, more specific, for example Xin et al. stated they intended their score to be used by practitioners at general practices in rural China (107).

Table 3.4 shows Google scholar identified that the number of citations for the articles in which these RATs were developed ranged from three to 131 with a median of 17.5. Examining these citations revealed that only three RATs (45,51,62) have had an impact study published. Heikes et al.'s RAT (62) was used as part of a community-based intervention which aimed to identity individuals at risk of a number of diseases including T2DM, this study found that

77% of individuals identified as at risk of diabetes were previously unaware they were at risk (120). Gray et al. (52) reported the use of the Leicester Practice Risk Score (LPRS) (45) in two studies (Let's Prevent Diabetes and Walking Away from Diabetes) to help identify individuals with NDH. In both rates of around 30% NDH or undiagnosed T2DM were found in those invited for an OGTT, which is a notable increase on the 20% seen in previous population-based OGTT screening in the same area (52). Khunti et al. (104) reported a cost per case analysis comparing screening blood tests alone to using a RAT, either the Leicester Self-Assessment (LSA) or LPRS (45,51), followed by screening blood tests. It found that using a RAT in combination with a blood test was more cost-effective than a blood test on its own, for both OGTT and HbA1c. For example, using HbA1c alone the estimated cost of detecting one case (NDH/T2DM) was £276, however this was lowered to £206 by screening using the LSA before the HbA1c screening and to £164 by using the LPRS first (104). Finally, Jones et al. (121) discusses the impact of using RATs in practice including the LSA. Noting that while the approach would work well in identifying individuals with NDH or undiagnosed T2DM, it would increase the workload in primary care.

Table 3.4 Number of citations of paper with risk assessment tools included in this systematic review as well as whether any external validations or impact studies were carried out for risk assessment tools in papers citing them

Risk assessment tool	Number of times cited	External Validation in citing paper	Impact study in citing paper
Barriga (108)	21	-	-
DuBose (109)	4	-	-
Handlos (105)	3	-	-
Heikes (62)	131	√	√
Hische (110)	3	-	-
Gray, 2010 (51)	29	√	√
Gray, 2012 (45)	17	√	√
Koopman (111)	18	-	-
Nelson (112)	26	√	-
Robinson (113)	16	-	-
Yu (106)	44	-	-
Xin (107)	11	-	-

Google scholar was used to carry out this search on 27th June 2014

Searching the citations identified in google scholar, four RATs (45,51,62,112) have had external validation reported in another paper. Two RATs, LSA and LPRS (45,51,62,112), were validated in a young South Asian population giving similar AUROCs to the internal datasets (115). Yet, when both RATs were validated in a Japanese cohort for the outcome of T2DM (either by FPG or HbA1c) they had considerably higher AUROCs of over 0.8 (116). When Nelson et al.'s RAT (112) was externally validated with T2DM, defined by FPG, as the outcome a sizably greater AUROC, 0.887, to the internal value, 0.74, was seen again (117). Phillips et al. (118) reports an external validation of Heikes et al.'s RAT (45,46,62,112), notably the AUROC with T2DM as the outcome, 0.70, is higher than the AUROC when NDH is the outcome, 0.67. Finally, Heikes et al.'s RAT was assessed in Korean populations for the outcome of T2DM, the AUROCs dropped noticeably compared to the internal value (119).

3.4 Discussion

This is the first systematic review with a search strategy that focuses on finding RATs that screen for individuals with NDH. The inclusion criteria set to find RATs that were applicable to general populations (hence the exclusion of RATs developed on pre-screened sample populations) in order to identify individuals who would benefit from interventions aimed at T2DM prevention.

This review has revealed that similar levels of internal performance can be seen when developing RATs using decision tree compared to logistic regression; with both Heikes et al. (62) and Xin et al. (107) seeing similar levels of performance for RATs developed using the two methods (both favoured the decision tree score due to its simplicity). The score developed by Yu (106) saw similar internal performance for the RATs developed using either SVM or logistic regression. Although in the majority of cases there was not a head-to-head comparison, the decision trees generally used fewer variables than the RATs developed using logistic regression. This may be useful because the logistic regression method is susceptible to over-fitting when there is a low EPV. However decision trees are prone to being unstable unless using specific statistical techniques (61) which develop the decision tree using several resampled versions of the data, as the RATs developed by Barriaga et al. used. This instability means a decision tree may not perform well in another dataset; this is seen with the three decision trees which have been externally validated, the AUROC for the external validation was at least 0.05 lower than the internal AUROC for each, with one having a drop in AUROC of over 0.1 for the external data compared to the internal data. In comparison the three RATs developed using logistic regression which were externally validated saw similar results.

In general the key characteristics of the data used to develop the RATs were well described; however many of the RATs yielded their sample from either advertising or opportunistic sampling rather than a population-based study. Careful consideration of the future use of the RAT needs to be taken before deciding if opportunistic sampling is appropriate, as the RAT will only be accurate for screening in similar settings. Yet, the intended future use of the RAT in several cases was an afterthought and did not inform the development

of the RAT. Not considering the intended use of the RAT may lead to poor decisions on methodological issues, such as the treatment of continuous variables. All but one RAT developed using logistic regression categorised all continuous data. This is not recommended (64), and the reason for doing so was only stated in two of the ten scores which were going to be used as paper-based scores; with six of the ten not reporting a specific use. Another key methodological issue of importance prior to developing RATs, which should be clearly detailed, is the treatment of missing data (40), which was not reported for 44% of RATs. None of the RATs used multiple imputation which has been advocated, instead of the simpler methods used in developing the RATs in this review, due to findings it leads to more valid results and better discrimination (65).

For the RATs developed using logistic regression, the most common way of selecting variables to be included was by backward elimination or another basic stepwise procedure. It has been proposed that stepwise procedures may miss sets of variables that fit well (122). Furthermore as stated in section 1.4.3, these methods have been criticised as they are prone to over-fitting (66,67). However over-fitting is not evident in the results of the external validations carried out for the logistic regression RATs in this review, this is likely due to good levels of EPV in their internal dataset. Variables should be selected using statistical methods alongside expert clinical knowledge/previous evidence (66,69). Expert knowledge/previous evidence were only considered when selecting the variables for a small number of the logistic regression risk scores (51,111,112). It is important that whichever method is used gives sufficient detail for it to be repeated, with the methodological decisions justified, as several fell short on one of these two areas.

Validation (either internal or external) of the risk scores, by evaluating the discrimination and calibration, assesses whether they work and thus should always be reported, with external validation being the gold standard (40). All RATs had some form of validation carried out on them with the majority performing bootstrapping, a resampling method that gives a good indication of how over optimistic the RAT may be (123). However, several RATs used randomly selected ‘training’ and ‘test’ datasets for their development and

validation, an approach which is likely to lead to overly optimistic results (75). External validation should be carried out before a RAT is considered for use in the real world (40); however this was only the case for seven (39%) of the RATs. All of the four RATs that had external validations carried out in another paper were for the outcome of T2DM. Interestingly three of four of these validations showed considerably better discrimination for the outcome of T2DM than had been seen internal. Only three (17%) of the RATs reported calibration (45,51,62), this is of concern as it is a vital characteristic to assess, particularly in the decision tree and SVM scores as these methods are prone to being unstable.

Evaluating the impact that a RAT has in clinical practice is a vital step that needs to be undertaken before any RAT can be advocated (40). Only three of the RATs (45,51,62) presented have had subsequent impact studies published, further highlighting the need for greater focus on the use of the RATs before development.

3.5 Conclusion and implications for thesis

This systematic review summarises 18 RATs which detect those at risk of NDH. Many of the findings emphasise those from other systematic reviews in the area discussed in section 1.4.3 (54-59). In general, greater thought to the intended use of RATs is needed before their development, as this will assist in making sensible decisions on how to develop the RAT, for example how to treat missing and continuous data. Other methodological issues of concern are the common lack of focus on calibration, external validation and impact studies of the RATs. Before these are used in practice, the level of calibration and validity of the RATs in the population of interest should be assessed.

One new finding is that the SVM method has been applied effectively to produce a RAT in this field. Though this technique needs further scrutiny as no external validation was carried out and it will therefore be examined in more detail, being included in the methods which are studied in the next chapter. The decision tree RATs included in this review which had external validations carried out displayed issues with over-fitting and therefore statistical techniques which have been advocated to deal with over-fitting are explored in the next chapter. The logistic regression RATs which had external validations carried out did not encounter problems with over-fitting despite using techniques some have criticised for being prone to this. This demonstrates that the more complex statistical techniques such as Least Absolute Shrinkage and Selection Operator (LASSO) and Least Angle Regression (LAR) which have been advocated to avoid over-fitting (68), may not be required when the number of EPV is adequate as has been found in simulation studies (123).

Chapter 4: Methods for developing risk assessment tools using cross-sectional data and a resampling study on the effects of the sample size of the development dataset on performance

4.1 Chapter outline

This chapter compares different methods for developing a risk assessment tool (RAT) for the outcome of non-diabetic hyperglycaemia (NDH) or undiagnosed Type 2 diabetes mellitus (T2DM) in a cross-sectional dataset with a view to establishing the best method, both statistically and practically. The methods compared are logistic regression, decision tree and support vector machine (SVM). In addition, this chapter includes a resampling study to assess the effects of differing the sample size of the development dataset on the performance of each of the methods.

4.2 Introduction to empirical comparison of methods

Systematic reviews of the area have shown that logistic regression is the most common method used to develop RATs which have been published for diabetes related prevalent outcomes (55,57,59,99). Other methods which papers included in these four systematic reviews utilised to develop a RAT for a prevalent outcome were decision trees and SVMs. While comparisons were carried out in a few of the papers found in the systematic review in Chapter 3, each paper only compared one method to logistic regression method rather than a comparison of all three methods (62,106,107). Furthermore, the two papers which compared decision trees to logistic regression (62,107) did not consider implementing extensions of the method such as boosting or bagging which have been shown to outperform the basic method in other types of datasets (124).

4.2.1 Findings of previous empirical comparisons in the medical field

Empirical comparisons of methods which model the risk of a binary medical outcome in cross-sectional data have been carried out (124-128); however there are several issues with generalising the results of these studies to establish which of the three methods considered in this chapter best discriminates blood glucose status. Firstly, only one of these papers (128) contains a comparison for a prevalent binary blood glucose outcome, the outcome of interest of this chapter. The performance of the method varies depending upon the setting (124), even between different medical areas, and thus it is not recommended to apply the results of such studies to outcomes in different medical fields (125). External validations of the performances were not included in any of these empirical comparisons, with internal validations being relied upon instead; despite external validation being the gold standard for verifying stability of RATs (40). Table 4.1 summarises the findings of these studies as well as detailing the study specific limitations in answering the question posed in this chapter. This will enable sensible decisions to be made to avoid the same shortcomings in this empirical comparison, as well as ensuring promising extensions of methods are investigated.

Caruana et al. carried out an empirical comparison of several methods in two unspecified medical datasets, as well as nine other non-medical datasets (124).

They found that the bagged trees and random forest methods had the best performance in one of the medical datasets; while random forest, logistic regression and neutral networks yielded the greatest levels of statistical performance in the other medical dataset. One concern with applying the results of this paper to the medical field in general is that the performance has been displayed as an average over eight metrics which are used in a variety of settings and thus no results for the preferred measure in medical field, area under the receiver operator curve (AUROC), are displayed (55,57,59,99). This study highlights the potential benefit of extensions to the basic decision tree method; therefore the empirical comparison in this chapter will investigate extensions when considering the decision tree method.

Cooper et al. compared several methods to discriminate for the outcome of mortality of hospital patients presenting with pneumonia (125). The paper finds that the error rates for the different methods are close and a much larger dataset would be required to detect statistical significance in error rates between the models. This study did not include the SVM method and the decision tree which was included was a hierarchical mixture with logistic regression, thus the two alternatives to logistic regression that have been applied in the setting of interest of this thesis were not included. Another limitation was AUROC was not reported, this is due to the different context, with incorrectly saying a patient will not have the event resulting in death meaning error rates were used.

Lehmann et al. evaluated numerous methods for discriminating between patients with Alzheimer's disease and healthy controls (126). They found that the more computationally intensive techniques, such as SVM and random forests, only performed a little better than traditional methods, like logistic regression. However this finding may be due to the small dataset used in this study (n=242).

Maroco et al. tested the performance of methods in discriminating mild cognitive impairment from Dementia in patients with a previous diagnosis of mild cognitive impairment (127). They found that SVM yielded the best AUROC yet it produced poor sensitivity. They also found the AUROC produced under logistic

regression and random forests were significantly better than the basic decision tree method. This study did not consider weighting the cases in order to try to improve the sensitivity of the SVM. Additionally they did not consider the extensions of bagging or boosting the decision tree method. Although the fact they showed another extension of the method, random forest, outperforming the basic technique; advocates extensions of the decision tree method being included in the comparison in this chapter.

Finally Tapak et al. carried out an empirical comparison of methods in discriminating diabetes status in Iranian participants (128). SVM performed extremely well in terms of AUROC and well in terms of sensitivity. Other methods, including logistic regression and random forest, resulted in poor sensitivity. Weighting of cases, which may have improved sensitivities, was not attempted in this study. Another limitation of this study is it did not include the decision tree method, despite the method having been used to develop several RATs for the outcome of interest and extensions of this method having performed well in discriminating other types of binary datasets (124).

Table 4.1 Summary of studies with empirical comparison of methods for discriminating binary outcome using medical dataset

Paper (first author and reference)	Dataset	Methods compared	Summary of findings	Limitations/weaknesses of study
Caruana (124)	Two unspecified binary medical datasets were included amongst 11 datasets from different settings. No further details about these datasets were given.	SVMs, neural nets, logistic regression, naive Bayes, memory-based learning, random forests, decision trees (including bagged trees, boosted trees and boosted stumps) It also examined the effect of that calibrating the models via Platt Scaling and Isotonic Regression.	Not one universally best method. Bagged trees and random forests performed well on one of the medical datasets. On the other, the best models were random forests, neural nets and logistic regression. The only models that never exhibit excellent performance on any problem are naive Bayes and memory-based learning.	The 2 medical datasets are for unspecified diseases. Only internal validation carried out. AUROC, the preferred measure in medical field, is not detailed.
Cooper (125)	Mortality of 14,199 hospital patients presenting with pneumonia based on findings at initial presentation. Methods were trained on 70% of dataset and tested on other 30%.	Neural network, rule-based model, hierarchical mixtures of experts (using decision tree structure), simple Bayesian model, Bayesian networks, logistic regression and K-nearest neighbour method.	'Most of the models error rates are sufficiently close such that a very large test database would be needed to reliably establish whether the rates differ statistically.'	Not for the outcome of interest. Due to context error rates were used rather than AUROC. SVM not considered. Decision tree not considered in standard form and extensions such as bagging, boosting and random forest were not considered.
Lehmann (126)	242 individuals from New York and Stockholm (116 patients with mild AD and 81 patients with moderate AD and 45 healthy age-matched controls). Two sets of	Linear discriminant analysis (principal competent and partial least squares, logistic regression (principal competent and partial least squares), bagged	10-fold cross-validation showed that modern computer-intensive methods (such as random forests, SVM and neural networks) performed only slightly	Small dataset. Not for the outcome of interest. Internal validation only.

	models were built, one aiming to discriminate between mild AD patients and healthy controls, the other set of models trying to discriminate between moderate AD patients and healthy controls.	decision tree, random forest, SVM and neural networks.	better than the traditional methods.	
Maroco (127)	Cohort study of 400 patients with initial diagnoses of mild cognitive impairment that were classified at follow-up as having either mild cognitive impairment or Dementia.	Linear discriminant analysis, Quadratic discriminant analysis, logistic regression, neural networks, SVM, decision trees and random forests.	Although SVM had the highest AUROC, it had the lowest sensitivity. Random forests and linear discriminant analysis were the best classifiers 'when taking into account sensitivity, specificity and overall classification.'	Not for the outcome of interest. Internal validation only. No weighting to attempt to address for poor sensitivity. Extensions of boosting and bagging decision trees not included.
Tapak (128)	6,500 Iranian subjects had diabetes status classified by fasting blood glucose.	Logistic regression, linear discriminant analysis, neural networks, SVM, fuzzy c-mean and random forests.	SVM has the best performance, performing very well for the AUROC and well for the sensitivity. The AUROCs for the other classifiers are good however the sensitivities, which are of great importance, are poor.	Internal validation only. No weighting to attempt to address poor sensitivity of methods was attempted. Extensions of boosting and bagging decision trees not included.

4.2.2 Datasets and methods compared

There are several novel methods which are not being included in the comparison in this chapter, such as neural nets or naive Bayes. The reason for this is that these methods are rarely employed by those in the medical field, as shown by the systematic reviews (55,57,59,99). Including fewer methods allows their extensions to be considered and thus an in-depth comparison of methods which those in this field are familiar with and most likely to use in the future to be carried out.

This chapter compares logistic regression, decision tree (including the extensions of bagged, boosted and random forest) and SVM (with both linear and radial kernels). The datasets used to evaluate these methods were summarised in Chapter 2. The ADDITION-Leicester dataset restricted to the 40-75 year olds was used to develop the RATs; while the Screening Those at Risk (STAR) dataset restricted to the 40-75 year olds was used to assess the external validity of the RATs. ADDITION-Leicester contains 6,390 individuals within this age range, while STAR contains 3,173 individuals within this age range. The outcome for this empirical comparison of methods was impaired glucose regulation (IGR) or undiagnosed T2DM.

As there is the issue of missing data in ADDITION-Leicester, this was dealt with using the recommended method of multiple imputation, which is detailed in section 4.3 (129). Additionally, as a sensitivity analysis the methods were applied using cases with complete data for the outcome and the seven predictors included in the Leicester Self-Assessment (LSA) score. Each method is firstly outlined and then results are presented for the outcome of IGR or undiagnosed T2DM. After each method has been employed to develop a RAT a discussion of the advantages and disadvantages of all methods both statistically and practically is given and the preferred method decided upon. This work informs the method chosen to develop the RAT with NDH or undiagnosed T2DM defined by glycated haemoglobin A1c (HbA1c) as the outcome in Chapter 6.

4.3 Multiple imputation of missing data

4.3.1 Concept

As shown in Table 4.2, the data being used for this empirical comparison has the issue of missing values. Only 67.5% of individuals have complete data on all candidate variables as well as the outcome variable. Carrying out the analysis using individuals with complete data on all candidate variables and the outcome variable would lead to losing around a third of the dataset and thus an undesirable reduction in statistical power (130). Furthermore, as many authors have pointed out a complete-case analysis often leads to biased results (70,131-135). Imputation methods use the observed values to estimate the missing values leading to reduced loss of information and if carried out carefully a reduction in the bias caused by the unobserved data (136). The most robust of the imputation methods is multiple imputation, which assigns several values for each missing value resulting in several copies of the dataset (130,137). These imputed values are based on the observed values of the other variables for that case and use a model based on the observed relationship between the variables across the whole dataset. Of course there is uncertainty in the relationship between the variables and this is well reflected by the variation in the numerous imputations for one missing value. The various versions of the dataset can then be analysed using complete-case methods with the results being combined by Rubin's rules (138). Multiple imputation assumes that data that is missing is missing at random (MAR), this means that the unobserved data is missing only due to the observed values.

Table 4.2 Summary Statistics of outcome and candidate variables in ADDITION-Leicester dataset

Variable	Observed data only summary	Observed and imputed data summary	Number of Missing values (% in brackets)
IGR or undiagnosed T2DM (%)	19.6	19.6	30 (0.47)
Age, years	57.3 (9.60)	57.3 (9.60)	1 (0.0)
Sex, Male (%)	47.7	47.7	0 (0)
Ethnicity, White European (%)	75.8	75.4	103 (1.6)
BMI (kg/m^2)	28.1 (4.99)	28.2 (4.99)	221 (3.5)
Waist (cm)	94.2 (13.1)	94.2 (13.0)	225 (3.5)
Current smoker (%)	14.5	14.4	237 (3.7)
Used high blood pressure drugs (%)	23.4	24.8	1,232 (19.3)
Previous high blood glucose (%)	10.5	11.1	1,101 (17.2)
Previous stroke (%)	2.1	3.2	1,646 (25.8)
History of high cholesterol (%)	17.4	19.1	1,530 (23.9)
History of high blood pressure (%)	27.8	29.3	1,438 (22.5)
History of Angina (%)	4.8	6.5	1,657 (25.9)
1st Degree Relative with diabetes (%)	25.2	26.6	1,204 (18.8)
Females with history of gestational diabetes (%)	1.3	1.3	0 (0)
Females with PCOS (%)	0.5	0.5	0 (0)
On steroids (%)	5.1	5.1	0 (0)

Values are mean (sd), unless stated

4.3.2 Potential pitfalls

One potential hazard of using multiple imputation is that data may be missing not at random rather than under the missing at random assumption (136); that is to say the reason it is missing is to do with its value, for example people with higher wages may not wish to give the value of their wage on a questionnaire. To deal with the issue that imputing some of the variables could lead to bias results, a sensitivity analysis was carried out using only the data with complete values for the outcome variable and the seven variables included in the LSA score, as stated in section 4.2.2. This will provide another comparison of the methods and if the results of this comparison differ greatly it will be attempted to understand why, a recommended precautionary checking process (136).

Another issue which needs careful consideration is which of the variables may explain the missing value, if variables predict either the value of a missing variable or the chance its missing it should be included in the model for its imputation even if it will not be in the subsequent analysis (139). In the multiple imputation here the systolic and diastolic blood pressure, which are not candidate variables for the RATs being built as they are not routinely known by members of the public and thus cannot be included in a self-assessment RAT, were included in the imputation of variables such as ‘used high blood pressure drugs’. The outcome variable will also be used in imputing missing values as not using it wrongly reduces the association between the variables and the outcome (140).

4.3.3 Details of multiple imputation carried out

In the dataset used here there are several variables with missing values, multivariate normal imputation and fully conditional specification (FCS) are two methods which are available in several statistics programmes to deal with a complex missing pattern such as this (141). FCS was used for the multiple imputation in this chapter, as there are lots of binary variables and this method allows them to be imputed under a logistic regression model rather than under the assumption of normality that would clearly be inappropriate (142,143). It also allows imputation under linear regression for variables with missing values which are continuous.

FCS, firstly gives all missing values a temporary value by randomly sampling with replacement from the observed values for that variable (141). It then replaces missing values for the first variable by sampling them from its conditional distribution, which is based on the relationship of the first variable with all the other variables included in its imputation model (144). This is repeated for all the other variables with missing values and once completed is called a cycle, after several cycles, say 10, the results are stable and a single imputation of the dataset has taken place. This procedure is repeated n times to give n imputed data sets. The FCS used here gave each variable with missing values a conditional distribution based on all the candidate variables, the outcome variable and the blood pressure measurements discussed above.

As there is a high percentage of incomplete data, simulation studies suggest between 20 and 40 imputations may be needed; 20 imputations were used here due to computation issues with the subsequent analyses with higher numbers of imputations (145). As non-normality of variables has been shown to adversely affect FCS imputations, continuous variables which did not follow a normal distribution were transformed before imputation and then transformed back along with the imputed values to their original scale afterwards (141). This imputation was carried out in Stata 13 (146) using the *ice* program (147).

The variables prescribed high blood pressure drugs and history of high blood pressure were combined into one hypertension variable after imputation, with a positive response for either variable being classified as hypertension. The STAR dataset has much lower levels of missing data, with 3,105 (97.9%) of individuals aged 40- 75 years old having values for all 15 candidate variables. Since using multiple imputation on the external validation assumes the assumptions made in using multiple imputation were correct, the external validation was carried out on complete-case data.

4.4 Comparison of methods

The methods are compared both in terms of both their statistical performance and in terms of the merits of implementing them in practice. The models' discrimination were measured using the AUROC, while the calibration was measured using the Brier score. It should be noted that the Brier score measures overall fit of a model, of which calibration is a component rather than being a measure of calibration alone. The calculations of these statistics are detailed in Chapter 2. The AUROC is a commonly used measure of discrimination in medical statistics (124). It is reported as a value between 0 and 1 or 0% and 100% and gives the probability that a randomly chosen case will have a higher value than a randomly chosen non-case. The 95% confidence intervals (CIs) are displayed for AUROCs calculated. The Brier score is a value between 0 and 1, the closer the value is to 0 the better the calibration of the model. Models' discrimination and calibration were calculated for both the internal and external data as this allowed an assessment of the external validity which is the gold standard for prediction models (40). 10x10% cross-validation was used to assess the internal performance of the RATs developed; except for the RATs developed using logistic regression which were assessed using 10-fold cross-validation as their development involved human input and thus the repetition was unfeasible (131).

The practical comparison of methods includes the setting in which the RAT could be used in practice, for example an app may be required. It also considers whether individuals will be educated about risk factors by completing the RAT or whether it is a 'black hole' which just tells them their risk status but not the reasons for this status.

4.5 Logistic regression

4.5.1 Method

Logistic regression is the most commonly used method for developing cross-sectional RATs with a binary outcome in the medical field (54,55,57,59,99), this is the main reason for its inclusion in this empirical comparison. The logistic regression model uses the values of the i[th] individual's explanatory variables ($x_{i1}, x_{i2}, \cdots, x_{ij}$) along with the coefficients ($\alpha,\ \beta_1, \beta_2, \cdots, \beta_j$) of the model to yield the log-odds of the i[th] individual having the outcome of interest (41). While under this equation the log-odds can take any value on the number line; transforming this value using the logistic function gives the probability of the i[th] individual having the outcome of interest, p_i, as a value between 0 and 1. Equation (4.1) gives the equation for the log-odds, while equation (4.2) gives the probability of an individual being a case given the values of their independent variables.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} \tag{4.1}$$

$$\mathrm{E}\left[Y_i \middle| x_{i1},\ x_{i2}, \cdots,\ x_{ij}\right] = p_i = \frac{e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}}}{1 + e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}}} \tag{4.2}$$

The logistic regression models for this analysis were fitted in Stata 13 (146). The values of the coefficients in the logistic regression model are assigned using the maximum likelihood method (41), which selects the values of the coefficients that are most likely given the data observed. As (4.3) displays, the outcome for each individual follows a Bernoulli distribution with parameter, p_i; meaning, as (4.4) shows, the probability of the outcome for individual i being a case is given by p_i while the probability of the outcome being a non-case is given by $1 - p_i$. (4.5) shows the equation that calculates the likelihood of the parameters ($\alpha,\ \beta_1, \beta_2, \cdots, \beta_j$) given the observed data. As can be seen it does this by multiplying the probably that the observed outcomes of interest for each individual would have happened given the parameters being considered and the observed independent co-variables for each individual.

$$y_i \mid X_{i1},\ X_{i2}, \cdots,\ X_{ij} \sim \text{Bernoulli}(p_i) \tag{4.3}$$

$$\left(Y = y_i \mid X_{i1},\ X_{i2}, \cdots,\ X_{ij}\right) = \begin{cases} p_i & \text{for } \ y_i = 1 \\ 1 - p_i & \text{for } \ y_i = 0 \end{cases} \tag{4.4}$$

$$\begin{aligned} \mathrm{L}(\alpha,\ \beta_1, \beta_2, \cdots, \beta_j) &= \prod_{i=1}^{n} \Pr\left(Y = y_i | X_{i1},\ X_{i2}, \cdots,\ X_{ij}\right) \\ &= \prod_{i=1}^{n} {p_i}^{y_i}\ (1 - p_i)^{(1 - y_i)} \end{aligned} \tag{4.5}$$

Stata 13 uses an algorithm to iteratively search for the most likely values of the parameters given the data, i.e. the values that maximise the likelihood function. The algorithm used in this analysis to fit the models is the Newton-Raphson method (148). Newton-Raphson begins with initial values for the parameters; it then adjusts these values to increase the likelihood at every iteration, stopping once the improvements in the likelihood are marginal.

Stata 13 uses the Wald's test to work out the p-values of the variables included in logistic regression models. Under the null hypothesis, that $\beta_j = 0$, the Wald's test statistic, W_j, is the square of the ratio of the coefficient for that parameter over the standard error for that parameter, as shown in (4.6). Under this null hypothesis, the test statistic, W_j, follows a Chi-squared distribution with one degree of freedom (149). This allows a p-value to be calculated for each variable in a model.

$$W_j = \left(\frac{\beta_j}{SE_{\beta_j}}\right)^2 \tag{4.6}$$

The variables to include in the logistic regression model selected here were initially chosen using backward elimination, the commonly favoured stepwise approach (131,132,150). Backward elimination starts with the full model, a model which contains all the candidate variables, it then removes the variable with the highest p-value and refits the model with all variables but this one. This process is repeated until all variables that are left in the model have a p-value lower than a pre-specified value, in this case 0.05.

Collins et al. points out that models being developed for use in clinical practice should not rely exclusively on statistical significance of variables considered (57). They advise that it may be sensible to keep variables known to be important risk predictors in the model even if they are not statistically significant for this particular dataset. Another important issue to consider is that self-assessment RATs used in clinical practice provide an educational message to those completing them about the relationship between the risk factors and the outcome. For this reason it may be wise to remove a significant variable from the model if its coefficient implies a relationship between the variable and the outcome which is not consistent with the known relationship or is contrary to general public health advice about that variable, for example if the model implies smoking is beneficial for the health outcome. Likewise a modifiable variable may be included which is not significant in order to educate those completing the RAT of the lifestyle changes they can make to reduce the risk of

the outcome, the Finnish Diabetes Risk Score (FINDRISC) included variables related to exercise and diet for this purpose (42).

For the reasons outlined in the previous paragraph, after backward elimination had been carried out it was considered whether any adjustments to the model should be made to take into account the health message and previous evidence as is recommended. Then interactions of the variables included as well as quadratic terms for the continuous variables were considered for inclusion as the relationship between a risk factor and the outcome may be more complex than a linear one. Due to the large number of interactions that likely would be tested, the interactions and quadratic terms were added to the linear terms if their p-value was lower than 0.01 and 0.05 respectively. This was done using a forward stepwise procedure, starting with the model with linear terms only. The RAT was yielded from this model by taking the expected value of y_i for an individual, given by (4.2), as their risk score.

As discussed in Chapter 1 the logistic regression can be used to create a simple RAT which an individual can use to calculate a score for their risk of the outcome using only pen and paper. This requires the continuous variables to be grouped and the coefficients of variables for groups to be rounded. Such a RAT was developed using the same variables as selected for the RAT suitable for an electronic platform, however no interaction or quadratic terms were added. This RAT allocated a score to an individual by giving them a score for each category they fall into, which was the coefficient for that category multiplied by 10 and rounded to the nearest whole number. The two RATs are later compared to the RATs produced by other methods with a discussion which takes into account both the performance and ways in which they can be implemented in practice.

In summary there are four stages to building the two RATs, with potentially a different model for each stage as follows:

1. Backward elimination starting with the full model of all candidate variables, using a significance level of $p<0.05$ for inclusion.
2. Model was then adjusted to take into account the health message of the model as well as previous evidence, as necessary.

3. Model which considers quadratic terms for continuous variables as well as interactions of all variables in the model at stage 2 was built. Adding these by forward selection starting with model built at stage 2, with statistical significance being $p<0.01$ for the interactions and $p<0.05$ for the quadratic terms. The model at this stage had an associated RAT calculated using equation (4.2) to work out the expected probability that an individual has the outcome given their independent variables. This RAT is appropriate for an electronic platform.
4. Model with categories for continuous variables, this model includes the same variables as model 2 and does not consider interactions or quadratic terms. For this model an associated RAT was calculated, this being the sum of the scores for all categories an individual falls into. The score for each category was the coefficient for that category multiplied by 10 and rounded to the nearest whole number.

4.5.2 Results

Performing stepwise backward elimination on the 15 candidate variables listed in section 4.3 results in the logistic regression model detailed in Table 4.3. This automatic approach yields seven variables, most of which have a relationship with the outcome which is the same as reported in the literature for T2DM and for other health outcomes. However being a current smoker decreases the risk of the outcome, this effect of smoking is not expected and with this being a RAT for self-assessment the impact of removing this variable on the statistical performance should be tested. Therefore a second logistic regression model was fitted with the smoking variable removed; this model also included body mass index (BMI) as this was only just not included in the automatically selected model and thus it may improve discrimination as well as giving an important health message. This logistic regression model is displayed in Table 4.4.

Table 4.3 Logistic regression model selected by automatic backwards elimination

Variable		Coefficient	95% CI	P Value
Age (years)		0.0439	0.036, 0.052	<0.001
Current smoker	No	Reference group		
	Yes	-0.256	-0.47, -0.044	0.018
Ethnicity	White	Reference group		
	Other	0.653	0.49, 0.82	<0.001
History of high blood	No	Reference group		
glucose	Yes	0.453	0.23, 0.67	<0.001
Hypertension	No	Reference group		
	Yes	0.333	0.17, 0.50	<0.001
First degree family	No	Reference group		
history of T2DM	Yes	0.502	0.33, 0.67	<0.001
Waist circumference		0.0322	0.027, 0.038	<0.001
(cm)				

Table 4.4 Logistic regression selected considering the health messages and previous evidence

Variable		Coefficient	95% CI	P Value
Age (years)		0.0459	0.038, 0.054	<0.001
BMI (kg/m^2)		0.0231	0.0027, 0.044	0.027
Ethnicity	White	Reference group		
	Other	0.677	0.51, 0.84	<0.001
First degree family	No	Reference group		
history of T2DM	Yes	0.493	0.32, 0.67	<0.001
History of high blood	No	Reference group		
glucose	Yes	0.439	0.22, 0.66	<0.001
Hypertension	No	Reference group		
	Yes	0.331	0.17, 0.50	<0.001
Waist circumference		0.0251	0.017, 0.033	<0.001
(cm)				

Interactions and quadratic terms of the variables included in the logistic regression model displayed in Table 4.4 were considered. However, no interactions or quadratic terms were found to be significant at the previously stated levels; hence the first RAT, the one appropriate for an electronic

platform, was based on the model displayed in Table 4.4. The associated risk score being the expected outcome given the known risk factors (variables); this is yielded using the coefficients from the equation given in (4.2). The resulting equation for the RAT is displayed in (4.7); this RAT clearly would require the aid of an electronic device to be completed by members of the public or in a healthcare setting.

$$logit^{-1}\left(\begin{array}{c} -7.68 \\ +\ (0.0459 \times age) \\ +(0.0231 \times bmi) \\ +0.677\ (if\ ethnicity\ is\ not\ white\ european) \\ +\ 0.493\ (if\ 1st\ degree\ family\ history) \\ +\ 0.439\ (if\ history\ of\ high\ blood\ glucose) \\ +\ 0.331\ (if\ hypertension) + \\ (0.0251 \times waist) \end{array}\right) \tag{4.7}$$

Grouping the continuous variables included in the second model detailed in Table 4.4 and using these variables along with the others included in that model allows a RAT which can be completed by hand to be produced. This is done as described in section 4.5.1; the score for being in each category of the variables is shown in Table 4.5 along with the logistic regression model from which it is derived. An individual's risk score is calculated by adding the score for each category they are in.

Table 4.5 Logistic regression model and scoring system for risk assessment tool with grouped continuous variables

Variable	Grouping	Coefficient	95% CI	P Value	Scoring
Age (years)	40-49	Reference group			0
	50-59	0.440	0.25, 0.63	<0.001	4
	60-69	0.861	0.66, 1.06	<0.001	9
	70+	1.16	0.93, 1.40	<0.001	12
History of high blood glucose	No	Reference group			0
	Yes	0.437	0.22, 0.66	<0.001	4
Ethnicity	White	Reference group			0
	Other	0.631	0.47, 0.79	<0.001	6
First degree family history of T2DM	No	Reference group			0
	Yes	0.475	0.30, 0.65	<0.001	5
Waist circumference (cm)	<90	Reference group			0
	90-99	0.500	0.29, 0.71	<0.001	5
	100-109	0.641	0.39, 0.90	<0.001	6
	>109	0.956	0.65, 0.79	<0.001	10
BMI (kg/m^2)	<25	Reference group			0
	25-29	0.137	-0.078, 0.35	0.212	1
	30-34	0.247	-0.017, 0.51	0.066	2
	≥35	0.458	0.127, 0.788	0.007	5
Hypertension	No	Reference group			0
	Yes	0.361	0.20, 0.53	<0.001	4

As can be seen in Table 4.6, the AUROCs of both the logistic regression RATs built in this section were good, in both the internal and external datasets. The Brier scores were almost identical for the internal and external dataset, whereas the AUROC increased slightly in the external dataset for the tool appropriate for an electronic platform and decreased a little for the tool with grouped continuous variables.

Table 4.6 Discrimination and calibration of logistic regression risk assessment tools

Risk assessment tool	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
Appropriate for electronic platform	0.701 (0.655, 0.748)	0.714 (0.692, 0.736)	0.145 (0.158)	0.144 (0.158)
Continuous variables grouped	0.723 (0.685, 0.762)	0.702 (0.680, 0.724)	0.143 (0.158)	0.145 (0.158)

4.6 Decision Tree

Several cross-sectional RATs in the area of diabetes with a binary outcome have used the decision tree method to develop the RAT (55,57,59,99). Those RATs developed for use in clinical practice tend to use only the basic version of the decision tree method; yet extensions of the method such as boosting, bagging or random forests have produced promising results in other empirical comparisons and are therefore considered here (61,124). Firstly, the basic decision tree method are outlined and the results using the simple version of this technique given before extensions of the method are introduced in turn along with their results.

4.6.1 Basic method

4.6.1.1 Concept of decision tree method

The fictional example for the outcome of whether an individual is currently employed or not in Figure 4.1 is used to explain the concept of the decision tree method. Decision trees start with all the individuals with outcomes that are to be predicted in one big group and then sorts them into several groups using successions of splitting criteria based on the explanatory variables. The decision tree in Figure 4.1, starts with all the individuals in one group on the left side; the 100% indicating that all the data is in this group, the two decimal numbers, 0.299 and 0.601, specifying the fraction of that group which have an outcome of 0 (not currently employed) and 1 (currently employed) in the training data respectively. As can be seen in Figure 4.1 discrete explanatory variables, such as whether an individual is in full time education, can be used to separate the data; as can continuous variables, such as age, however continuous data require a single cut point to be chosen to split the data into two new groups. As can also be seen in the figure, once data have been split into different groups, these new groups can be split using different splitting criteria to one another. Also, note explanatory variables may be used in more than one splitting criteria in a decision tree, such as age is in the example in Figure 4.1.

In the cases where the data has a binary outcome, such as in this chapter, each group is assigned an outcome based on the proportion of the two outcomes observed in that group from the training data; in Figure 4.1, this assigned outcome for each group can be seen at the top of its box, being either 0 or 1. In

unweighted decision trees the outcome assigned to a group is just the majority outcome observed in that group for the training data; this is the case in the example in Figure 4.1. Predicting the outcome of an individual using a decision tree just requires the following of the 'branches' with the splitting criteria which match that individual's data until an end group is reached, one which is not split further, the individual is then predicted to have the outcome which has been assigned to this group. For example, using the decision tree in Figure 4.1 to predict whether a 20 year old who is not in full-time education is currently employed would lead to the prediction that they are currently employed. This is because being less than 21 years old they would follow the lower branch into the group with 20.8% of the data. From there they would again follow the lower branch, as they are not in full time education, leading them to be in the group with 7.8% of the data. As this group is not split any further the individual is predicted the outcome this group is labelled with, in this case 1 (or currently employed).

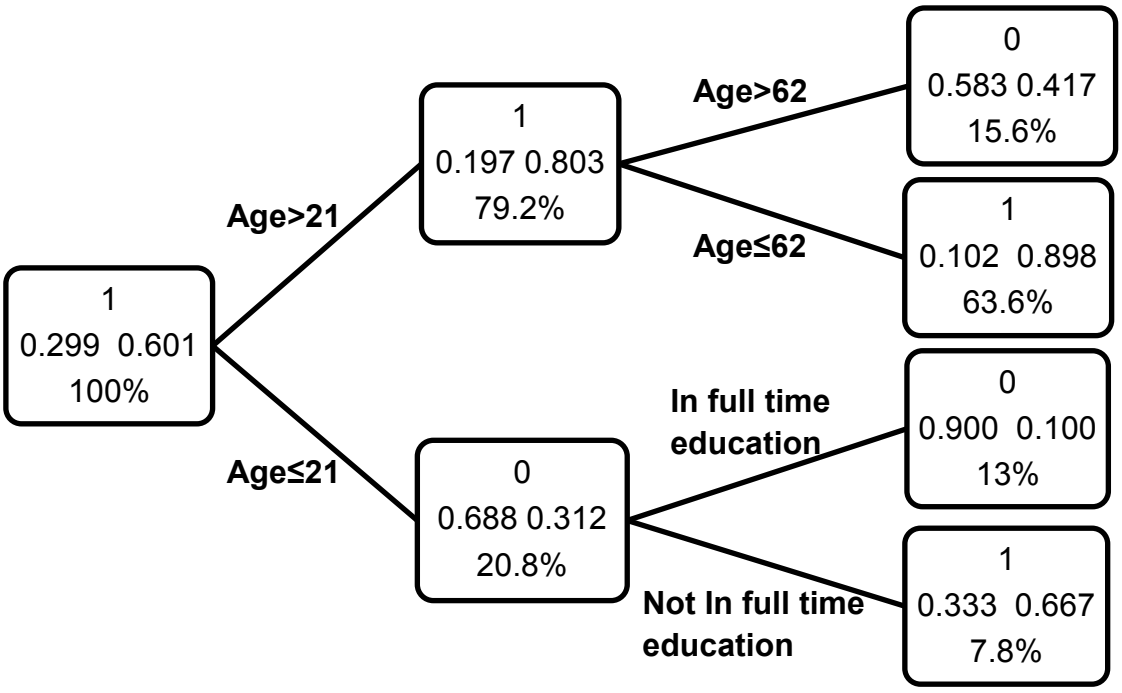


Figure 4.1 Fictional example of a decision tree for whether an individual is currently in employment

4.6.1.2 Details of method

Now the concept of this method has been outlined in the previous section, the method is described here in greater detail and using the recognised terms. Decision trees start with all individuals grouped together in the *root node*, at this point the only information known about an individual is that they are part of the dataset and therefore the chance they have the outcome of interest is the prevalence of that outcome in the dataset. The decision tree then separates the individuals into two subgroups, choosing the explanatory variable and cut-point that 'best' separates individuals into those with a high chance of having the outcome and those with a high chance of not having the outcome (60). This process of splitting the individuals is then repeated for the subgroup which results in the 'best' separation of the data, again considering all possible cut-points of all explanatory variables. This process of splitting the data is repeated until the discriminative benefit of adding more splits is smaller than the penalty chosen for making the tree more complex. Nodes are labelled as cases or non-cases depending on what results in the lowest amount of misclassification of outcomes from the training data in that node, when cases are unweighted this is simply the majority outcome in that node of the training data. The nodes at the end of the decision tree, known as the terminal nodes or leaf nodes, are used to predict the outcome of individuals based on their explanatory variables.

Most methods for choosing the best splits and when to stop splitting the tree are based on the impurity of the nodes. Impurity is a measure of how skewed the distribution of the outcome classes are in a node, lower impurity means a higher fraction of the outcome are from a single class and higher impurity means there are more even levels of the two classes of outcome (151). The basic decision trees in this thesis have been built in the statistical programme R using the *rpart* package (152), with the Gini index being used to measure impurity. The Gini impurity of a node gives the probability of wrongly guessing a randomly chosen outcome by randomly assigning a class to it based upon the fractions of the classes observed in that node (151). As can be seen in (4.8) the impurity for the node i, D_i, is calculated by summing the squared proportion of cases and the squared proportion of non-cases and subtracting this from one. Note that when

all individual in a node belong to the same class $D_i = 0$, this node is perfectly classified.

The impurity of a decision tree is the probability of wrongly guessing a randomly chosen outcome by randomly assigning a class to it based upon the fractions of the classes observed in the terminal node it belongs to. This means any proposed decision tree's impurity can be calculated by taking a weighted average of the impurities of its terminal nodes, the weighting , f_i, for each node being the fraction of the data in that node. The *rpart* function chooses the split at each step which reduces the impurity of the weighted terminal nodes the most. The cost-complexity criterion is used to decide what size tree should be selected; it balances the reduction in impurity of a larger decision tree with simplicity of a smaller decision tree. The cost-complexity, $CC(T)$, is calculated for the decision trees considered using equation (4.9), where T is the number of splits in the tree being considered and λ is the penalty term for having a more complex tree.

$$D_i = 1 - \sum_k p_{ik}^2 \tag{4.8}$$

$$CC(T) = \sum_{\substack{terminal \\ nodes}} f_i D_i + \lambda |T| \tag{4.9}$$

The *rpart* function continues to add the 'best' split at each iteration while $CC(T)$ is reduced, stopping once $CC(T)$ can no longer be decreased by doing so. Selecting the penalty term for complexity is difficult; choosing a value too small will result in overfitting of the decision tree, while picking a value which is too high will result in a model that does not fit the data very well. An advocated technique to decide a suitable size for the tree to be grown is cross-validation (153). Cross-validation splits the dataset into k subgroups; it builds k fully grown decision trees, each time using all the dataset apart from one of the subgroups. The misclassification error of the unused subgroup is calculated for every size of tree in each of the decision trees built, the size of tree that has the lowest misclassification error against all the validation subgroups is chosen as the optimal size to prune the decision tree to. A decision tree is then built on the whole dataset using the pruning parameters found by cross-validation. The

basic decision trees assessed here used a 10-fold cross-validation to choose the parameter that determines the size of the decision tree that was built.

Finally, several values were considered for weighting the importance of misclassifying a case compared to non-case, this was done for two reasons. Firstly, the prevalence of the outcome being 19.6% means a decision tree without weighting will have a very high specificity however this will likely be at the expense of having a very low sensitivity. Secondly, in this context, as is the case for many medical applications, a false negative is a worse error than a false positive (108). Especially since this is the first step in the screening process, meaning a false negative wrongly reassures individuals whereas a false positive is likely to be found to be incorrect at the next stage of the process. The extension of the impurity function displayed in (4.8) to include costs which account for whether a case or non-case has been misclassified is detailed elsewhere (154).

A paper included in the systematic review in the previous chapter which used weights in developing its RATs highlighted that there are not any clear guidelines about what weighting to use (108). The dataset in that paper had a similar prevalence of outcome to dataset used in this chapter. The previous paper opted to use a weighting of 8:1, and this yielded high sensitivities however specificities were low as a result. As it was unknown what weighting would lead to a satisfactory balance, a range of weightings were investigated. Weights of 1:1, 2.5:1, 5:1, 7.5:1 and 10:1 for the importance of misclassifying a case compared to non-case were assessed. This range goes from unweighting to slightly above the level used in the paper, as it has shown this is large enough to yield high levels of sensitivity. The minimum number of observations in a terminal node will be assessed for 1, 10 and 100 and the complexity parameter evaluated at 0.1, 0.01 and 0.001. Only the results with complexity parameter set to 0.01 and 10 as the minimum number of observations in a terminal node are displayed in this chapter.

4.6.1.3 Basic method results

Table 4.7 shows the internal and external AUROCs of decision trees with a variety of weights; the decision trees with weightings of 2.5:1 and 5:1 gave the best discrimination, although these AUROCs were still low, indicating poor discrimination. The Brier scores increase with increased weighting for cases. Changing the weights of the decision trees greatly affected the sensitivity and specificity; with sensitivity being very low for the unweighted tree and very high for the tree with a weighting of 10:1, conversely the specificity was very high for the unweighted tree and very low for the tree with a weighting of 10:1.

Table 4.7 Internal and external AUROCs and Brier scores of decision trees with various weighting for cases-to-non-cases

Weighting of cases compared to non-cases	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (outcome index variance)	**External Brier score** (outcome index variance)
1 :1	0.500 (0.500, 0.500)	0.500 (0.500, 0.500)	0.158 (0.158)	0.158 (0.158)
2.5 :1	0.630 (0.570, 0.690)	0.627 (0.603, 0.652)	0.182 (0.158)	0.187 (0.158)
5 :1	0.636 (0.580, 0.692)	0.637 (0.613, 0.660)	0.262 (0.158)	0.275 (0.158)
7.5 :1	0.592 (0.555, 0.628)	0.582 (0.565, 0.600)	0.341 (0.158)	0.346 (0.158)
10 :1	0.589 (0.552, 0.626)	0.584 (0.567, 0.602)	0.395 (0.158)	0.399 (0.158)

Decision trees with 10 cross-validations, 10 as the minimum number of observations in terminal node and complexity parameter=0.01

Table 4.8 shows the decision tree with a weight of 5:1 produced the most suitable balance of sensitivity and specificity for this context and thus this weighting will be used when building decision trees to assess extensions of the decision tree method. Similar results were produced under this method with the minimum number of observations in the terminal node set to 1 and 100. Decreasing the complexity parameter to 0.001 led to only a modest increase in AUROCs; with the greatest external AUROC being 0.66.

Table 4.8 Internal and external sensitivity and specificity of decision trees with various weighting for cases-to-non-cases

Weighting of cases compared to non-cases	**Internal cross-validated sensitivity** (95% CI)	**External sensitivity** (95% CI)	**Internal cross-validated specificity** (95% CI)	**External specificity** (95% CI)
1 :1	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
2.5 :1	26.9 (17.0, 36.7)	28.2 (24.8, 32.0)	88.8 (82.9, 94.8)	87.3 (86.0, 88.6)
5 :1	68.4 (53.5, 83.2)	74.1 (70.4, 77.4)	54.8 (39.5, 70.1)	49.0 (47.0, 50.9)
7.5 :1	91.3 (84.5, 98.0)	88.7 (85.9, 91.0)	22.1 (14.7, 29.4)	25.8 (24.1, 27.6)
10 :1	94.1 (89.1, 99.2)	93.4 (91.1, 95.2)	17.2 (10.5, 23.8)	18.5 (17.1, 20.1)

Decision trees with 10 cross-validations and 10 as the minimum number of observations in terminal node and complexity parameter=0.01

Figure 4.2 displays that six terminal leaves were included in the decision tree with a weighting of 5:1, four variables were included with waist being used in the splitting criteria at two nodes.

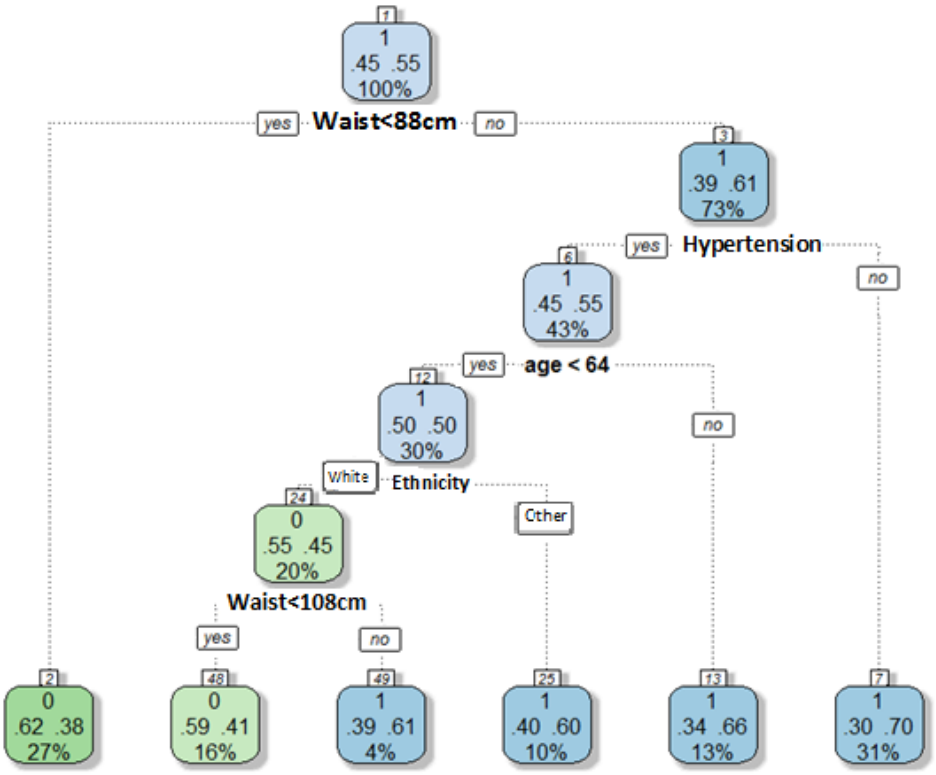


Figure 4.2 Decision Tree with case weights=5, cross-validations=10, minimum number of observations in terminal node=10 and complexity parameter=0.01

4.6.2 Boosted decision tree

Boosting is a method which aims to increase the accuracy of the decision tree method by iteratively building numerous trees with the weights for misclassified outcomes being increased, or boosted, after each tree is built (155). This means the later trees in the process place a greater emphasis on correctly predicting the outcomes which have most often been misclassified. Firstly, a fraction of the dataset is chosen without replacement to develop the first decision tree, at this point the usual weights are used as no outcomes have yet been misclassified. After the tree has been built the weights of the outcomes are adjusted, increasing the weights of outcomes that were misclassified and as a result decreasing the weights of those that were correctly classified. This process is then repeated with a new subsample chosen from the dataset being used to build a tree and the weights again being recalculated. After this process has

been repeated a number of times, say 1000, the probability an individual is a case can then be calculated, by averaging how often it was classified as a case. To avoid over-fitting the contribution of each tree to this average is reduced the further into the iterative process it was fitted, the parameter that controls this is known as the shrinkage of the learning rate (156).

An initial weight of 5:1 was used for the misclassification of the cases compared to non-cases as this was found to yield a suitable balance between sensitivity and specificity in section 4.6.1.3. As Hastie et al. suggest limiting the trees to have between 4 and 8 terminal nodes leads to sensible results which are not over-fitted, therefore maximum depths of 2 and 3 were assessed here (157). Friedman proposes the fraction of the dataset used for each subsample should be between 0.5 and 0.8 for datasets of moderate size such as this one, as smaller values avoid over-fitting 0.5 was initially considered here with various values subsequently assessed (158).

This work will be carried out using the package *gbm* in R (159). The shrinkage of the learning rate interacts with the number of trees used in determining the fit and thus whether over-fitting occurs (156). Therefore shrinkage will be assessed at a range of values at 0.01 and 0.0001 as well as its default, 0.001. 10-fold cross-validation will be utilised to select an appropriate number of trees to use.

4.6.2.1 Boosted method results

Cross-validation for the number of trees to include recommended the maximum number of trees possible regardless of the initial number of trees included in the model, meaning it was not computationally viable to build a model with the number of iterations the cross-validation results suggested were necessary. Although, as can be seen from Table 4.9, while the discrimination of the method rose when increasing the number of trees from 500 to 1,000 and again when increasing the number of trees to 5,000; this pattern did not continue for the model with 10,000 trees. This suggests that the AUROC will remain at around 0.70 even if the number of trees was increased further. The Brier scores display the same pattern. Similar results were also seen when the maximum depth was

set to 3, 0.75 was the fraction of the data used at each iteration or the shrinkage varied.

Table 4.9 Internal and external AUROC and Brier scores for boosted decision trees with varying number of iterations

Number of Trees	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
500	0.659 (0.633, 0.685)	0.664 (0.640, 0.687)	0.271 (0.158)	0.274 (0.158)
1,000	0.670 (0.644, 0.696)	0.678 (0.655, 0.701)	0.265 (0.158)	0.269 (0.158)
5,000	0.700 (0.674, 0.723)	0.703 (0.681, 0.725)	0.247 (0.158)	0.254 (0.158)
10,000	0.703 (0.677, 0.729)	0.703 (0.681, 0.725)	0.243 (0.158)	0.250 (0.158)

Boosted decision trees with 5:1 case-to-control weighting, shrinkage=0.001, fraction=0.5, maximum depth=2, minimum number of observations in terminal node=10, cross-validations=10

4.6.3 Bagged decision tree

Bagging builds a number of decision trees using several bootstrapped samples of the dataset (160). A bootstrapped sample is simply a random sample of the dataset with replacement. An individual's outcome can then be predicted under all the trees; with the outcome which it is classified as most frequently using all the trees being its predicted outcome under the bagged decision tree.

The *ipred* package in R was used to build the bagged decision trees considered in this section (161). 100 bootstrapped samples were used due to computational constraints, research showing in previous datasets that this has been ample to optimise the discrimination of this method (160). A sample size equal to the overall size of the dataset was used for each bootstrapped subsample, this results in roughly 63.2% of the dataset being included at least once (162). Various maximum depths of the trees (2, 5, 7 and 10) and complexity parameters (0.1, 0.01 and 0.001) were considered. The reason cross-validation was not carried out to choose these values as it is too computational intensive as the package has acknowledged (161). The minimum number of observations in a terminal node was assessed for 1, 10 and 100.

4.6.3.1 Bagged method results

Table 4.10 shows that although the discrimination improved when the depth of trees was increased from two to five; the AUROCs, both the internal cross-validated and external, were not increased further by allowing the maximum depth to increase to seven or ten, with the AUROCs being at 0.68 in both datasets. The same pattern was seen for the various complexity parameters, maximum depths and for various minimum sizes of terminal node. The Brier scores for the model with a maximum depth of two were slightly worse than the models with a maximum depth of five, seven and ten.

Table 4.10 Internal and external AUROC and Brier scores for bagged decision trees with varying maximum depths

Maximum depth of trees	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
2	0.645 (0.621, 0.677)	0.659 (0.635, 0.683)	0.152 (0.158)	0.151 (0.158)
5	0.680 (0.652, 0.707)	0.682 (0.659, 0.706)	0.148 (0.158)	0.148 (0.158)
7	0.678 (0.651, 0.706)	0.681 (0.658, 0.704)	0.149 (0.158)	0.149 (0.158)
10	0.677 (0.649, 0.704)	0.680 (0.657, 0.703)	0.149 (0.158)	0.149 (0.158)

Bagged decision trees with 5:1 case-to-control weighting, complexity parameter=0.001, bootstrapped samples=100, minimum node size=10

4.6.4 Random forest

Random forests like bagged decision trees use several bootstrapped samples of the dataset (163). The way random forests differ from bagged decision trees is they only consider a random subset of the explanatory variables to base each split on, this means there is more variation between trees (164). An individual's outcome is predicted under each of the decision trees, with the outcome it is classified as most frequently using all the decision trees being its predicted outcome under the random forest.

The *randomForest* package in R was used to build the random forests considered here (165). As with the bagged decision tree the size of the bootstrapped samples are chosen to be the same as the size of the dataset. The number of variables considered at each split was four, since there are 15 candidate variables and its proposed that the square root of the candidate variables should be the number of variables considered at each split (165). No penalty parameter for increasing the number of splits in a tree is included in the *randomForest* package. The minimum size of terminal node was varied with the default 1 being assessed along with 10 and 100. The number of trees used was also varied, with 25,100 and 1000 being assessed along with the default, 500.

4.6.4.1 Random forest results

As can be seen in Table 4.11 the random forests discriminate better in the external dataset compared to the internal dataset, the model with ten as the minimum size of terminal node gives the best AUROC in both the internal and external data. The Brier scores show the calibration became worse as the minimum size of the terminal node was increased. Similar results were seen with models with different numbers of trees.

Table 4.11 Internal and external AUROCs and Brier scores for random forests with varying the minimum numbers of observations in a terminal node

Minimum size of terminal node	Internal cross-validated AUROC (95% CI)	External AUROC (95% CI)	Internal cross-validated Brier score	External Brier score
1	0.635 (0.612, 0.658)	0.655 (0.631, 0.679)	0.284 (0.158)	0.221 (0.158)
10	0.655 (0.631, 0.678)	0.680 (0.656, 0.703)	0.446 (0.158)	0.370 (0.158)
100	0.638 (0.614, 0.661)	0.678 (0.655, 0.701)	0.638 (0.158)	0.657 (0.158)

Random forests with 5:1 case-to-control weighting, four variables considered at each split, trees=500

4.6.5 Summary of results

The basic decision tree method gave poor discrimination in the internal data, with similar levels of discrimination in the external dataset showing the models built were stable. The extension of boosted decision trees gave good discrimination once a large enough number of trees were used, 5,000 was sufficient. However, the Brier scores indicated the calibration of this method was poor with 5:1 case-to-non-case weighting. Bagged decision trees resulted in reasonable discrimination and calibration when the maximum depth allowed was at least five. Random forests produced poor calibration and the discrimination in the internal data was worse than that seen for the boosted or bagged decision trees. Therefore, the results from the comparison using multiple imputed data in model developed support the use of the boosted and bagged decision tree method.

4.7 Support Vector Machine

The systematic review in Chapter 3 found only one RAT which was developed using the SVM method (106). In the dataset used the SVM method provided comparable discrimination to the logistic regression method. In empirical comparisons of methods for discriminating binary outcomes in the medical field, discussed in section 4.2.1, the SVM has performed well with some particularly encouraging results in a couple of the studies (127,128). This section will firstly outline the concept of the SVM method, and then give a more detailed description of the linear and non-linear variations of the method that will be included in this comparison. Finally, the results of the two SVM models will be reported.

4.7.1 Concept of SVM method

A SVM aims to split the data into events and non-events using a multidimensional hyperplane based on the explanatory variables (106). A hyperplane is a subspace of one dimension less than the whole space being considered. For example, a line is a hyperplane of 2-dimensional space; and a 2-dimesional plane is a hyperplane of 3-dimesional space. In the illustrative examples used to explain the concept of SVM models in this section, only two explanatory variables will be used to allow the SVM models to be depicted easily as they will only be in 2-dimesions.

Figure 4.3 and Figure 4.4 show an illustrative example of a linear SVM using just two explanatory variables, with the outcome classes colour coded in red and green. The model utilises the position of the explanatory variables in relation to one another and aims to find a hyperplane that separates the two classes. In this case as there are only two explanatory variables the hyperplane is a line, furthermore as the SVM is linear the hyperplane is a straight line. As can be seen in Figure 4.3 there are numerous such hyperplanes that split the data cleanly into the two outcome classes. When this is the situation the optimal hyperplane is the one that has the biggest distance, known as margin, between itself and the support vectors. The support vectors are the data points that are closest to the hyperplane when measuring the distance of the data points from the hyperplane. The hyperplane must be chosen such that there is at least one support vector from each of the two classes, meaning there is an equal margin on each side of the hyperplane.

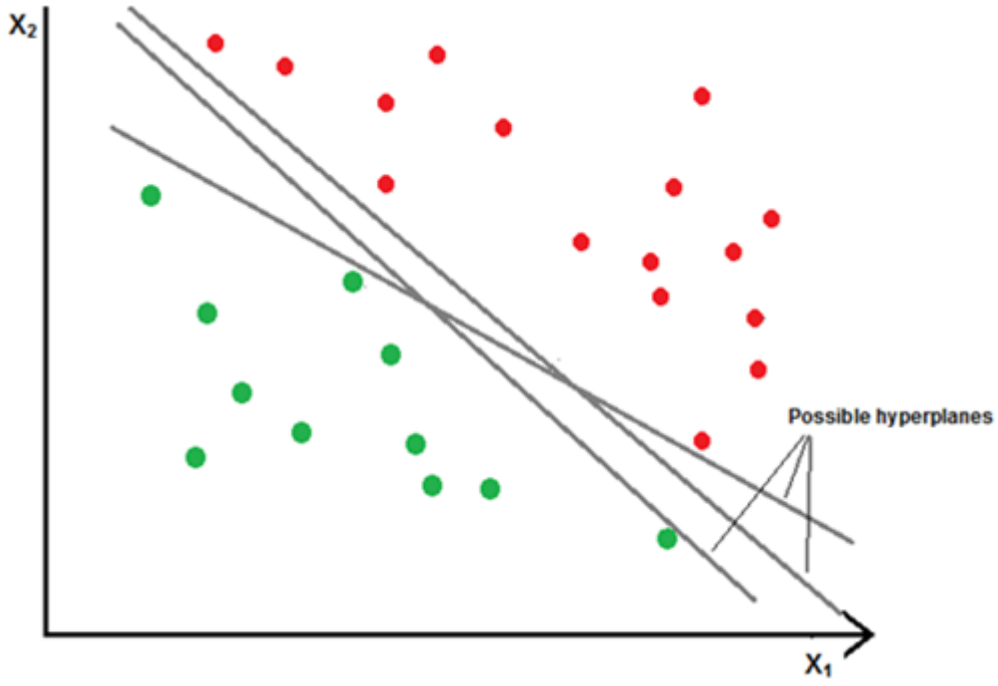


Figure 4.3 Possible hyperplanes of an illustrative example of linear SVM with hard margin using only 2 explanatory variables

Figure 4.4 shows the optimal hyperplane which in this example has been chosen based on the position of the five support vectors which lie at the edge of the margin, two green outcomes along one supporting hyperplane and three red outcomes along the other supporting hyperplane.

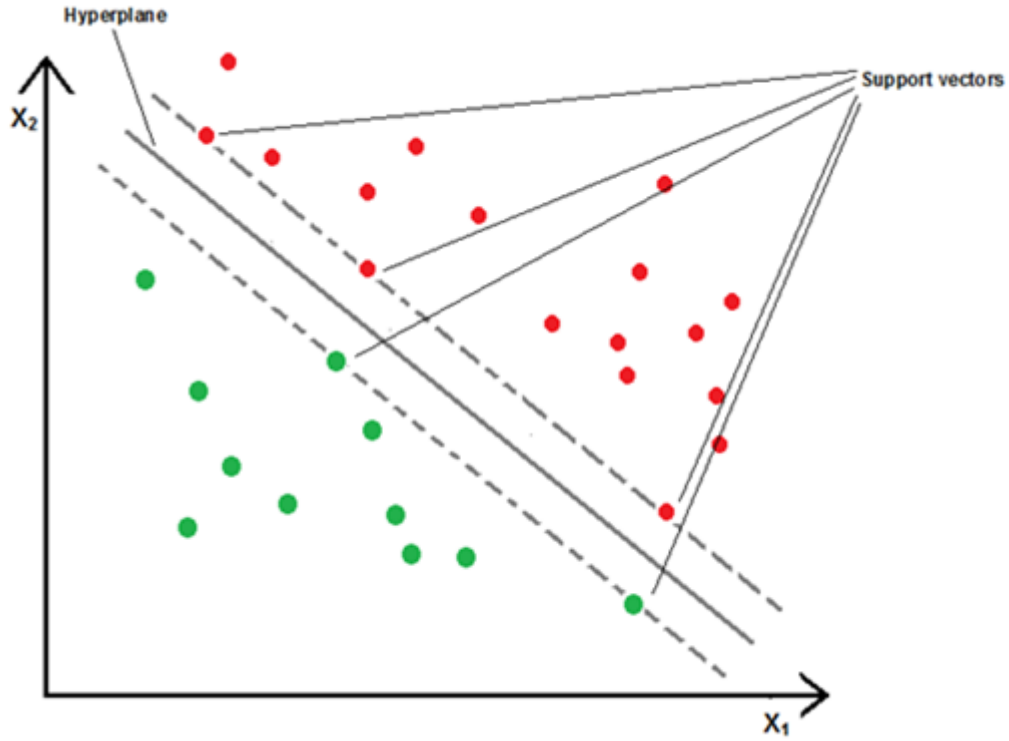


Figure 4.4 Illustrative example of linear SVM with hard margin using only 2 explanatory variables, with optimal hyperplane and support vectors shown

The example displayed in the Figure 4.3 and Figure 4.4, uses a 'hard margin' meaning no vector can be on the wrong side of its supporting hyperplane; as can be seen in Figure 4.4 there are no vectors within the margin or amongst the opposite outcomes. However, as is often the situation in medical datasets, sometimes the two classes cannot be completely separated by a hyperplane meaning that a SVM with 'hard margin' cannot be built. In order to overcome this limitation the method has to be adapted slightly, this was achieved by the introduction of 'soft margins' (166).

As shown in Figure 4.5, SVMs with soft margins allow vectors to be on the other side of their supporting hyperplane. How many vectors are on the opposite side of their supporting hyperplanes depends on the amount of ‘slack’ that is allowed in the SVM. In the example depicted in Figure 4.5, the three vectors with non-zero slack values are indicated by the bold lines, the slack in each case simply being the distance these vectors are from their supporting hyperplane. SVMs with a soft margin separate the rest of the vectors, those with zero slack, aiming to maximum the margin as before, however the possible hyperplanes are penalised for each vector with a non-zero slack proportional to the value of that slack. Thus as labelled in Figure 4.5 vectors that have non-zero slack as well as those along the supporting hyperplanes are support vectors meaning they determine the hyperplane chosen. The equation that has to be optimised when building a linear SVM with soft margin will be detailed in the next section.

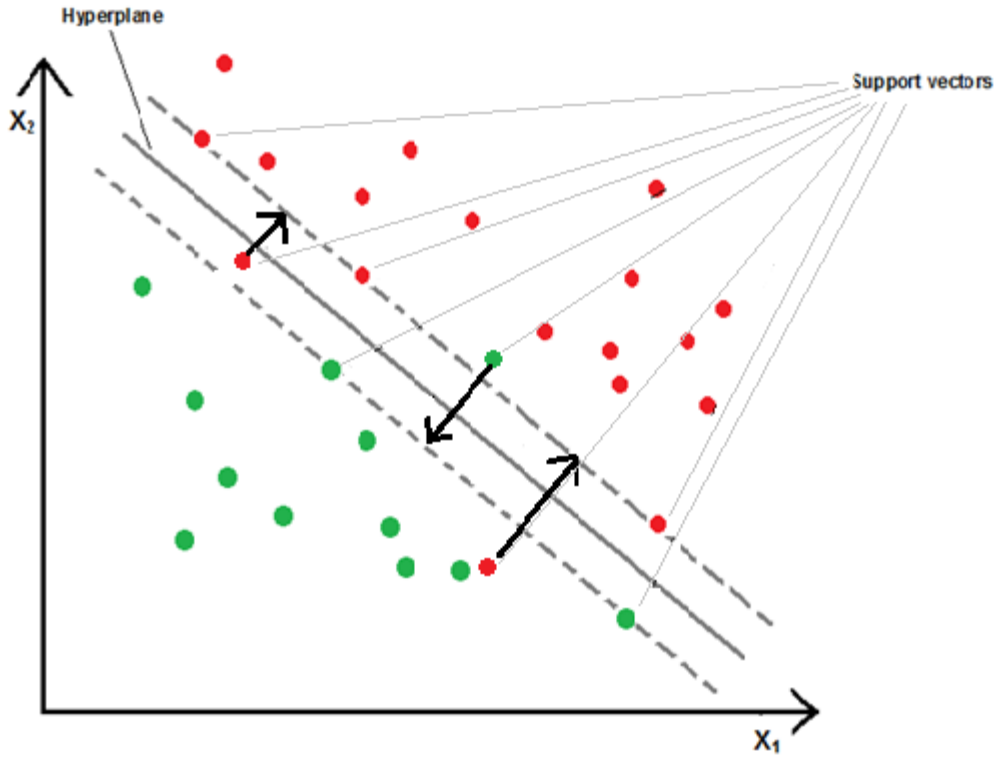


Figure 4.5 Illustrative example of linear SVM with soft margin using only 2 explanatory variables

The SVMs discussed so far in this section have been linear, however sometimes a linear SVM will be very ineffective in separating two classes and lead to lots of misclassification. Fortunately, numerous different shapes of hyperplane can be used to split the outcome classes by adjusting part of the equation to be optimised (106,167). This adjustment replaces part of the equation with a non-linear kernel function, thus the shape which is used for the hyperplane of a SVM is known as the type of kernel it has used. Figure 4.6 shows an example of data for which a linear SVM would have performed poorly, demonstrating the radial kernel used provides a much more suitable SVM in this case.

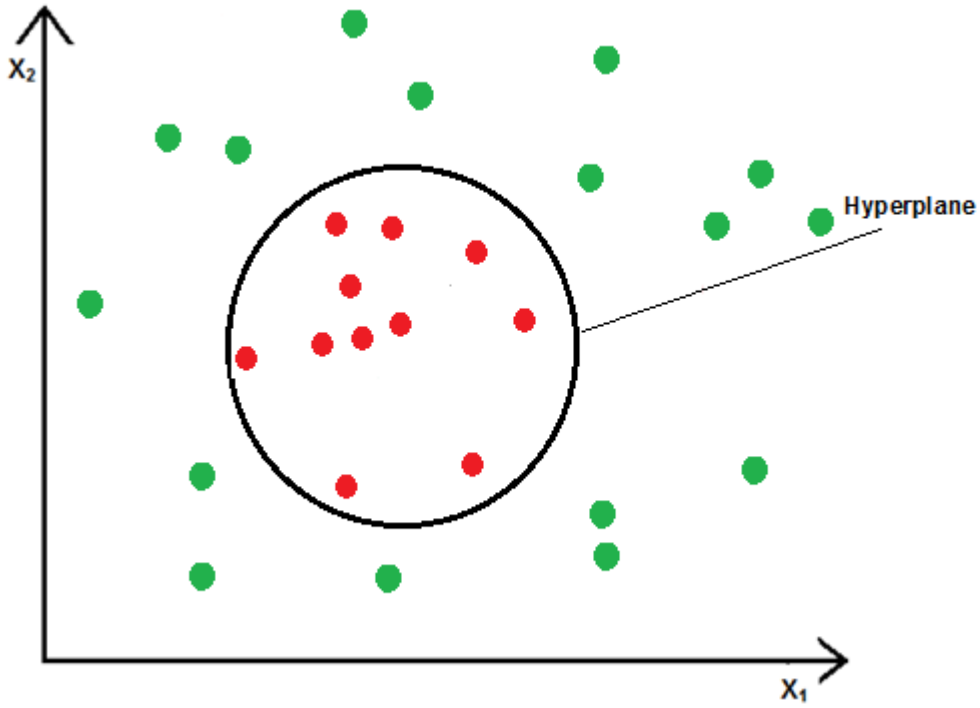


Figure 4.6 Illustrative example of radial SVM using only 2 explanatory variables

The comparison in this chapter includes a SVM with a linear kernel and a SVM with a radial kernel. The first reason for selecting these kernels was they are the most commonly used kernels for this type of classification problem (126-128). The second reason was when comparing the performance of four kernels (linear, polynomial, radial and Sigmoid), in discriminating two different diabetes-related binary outcomes; Yu et al. found that the SVM with linear kernel performed best for one outcome, and the radial kernel lead to the best performance in the other (106).

4.7.2 Details of method

This section gives a more in-depth explanation of the method, detailing the mathematical equations behind the concept. The equations relating to linear SVMs are given first, after which the way in which these equations are adapted for radial SVMs is detailed.

4.7.2.1 Details of linear SVM

Firstly, each outcome variable , y_i, is labelled either -1 if it is not the outcome of interest or 1 if it is the outcome of interest. If a linear SVM with hard margin can be implemented, this means there are multiple hyperplanes that cleanly separate individuals in to the two outcome classes. Hyperplanes can be notated as the set of explanatory variables $\boldsymbol{X}$ that satisfy (4.10); where $\boldsymbol{W}$ is the normal vector to the hyperplane and b is a constant (166,168). As the optimal hyperplane, $(\boldsymbol{W}^*, b^*)$, separates the data this means the inequality (4.11) is satisfied for all individuals in the dataset. Taking the modulus of the express on the left hand side of the function given in (4.10), $|\boldsymbol{W}.\boldsymbol{X}_i + b|$, gives the perpendicular distance each individual data point is from the hyperplane, $(\mathbf{W}, b)$. As explained in section 4.7.1, the optimal hyperplane is chosen to maximise the distance of the nearest data points to the hyperplane, the support vectors, from the hyperplane. As multiplying a hyperplane by a scalar, say k, results in the same hyperplane there are multiple solutions which are all the optimal hyperplane. This issue can be overcome by stipulating the normal vector of the optimal hyperplane , $\boldsymbol{W}^*$, must have a norm, the dot product of the vector with itself, of a specific value and then finding the maximum margin of such hyperplanes. This is equivalent to setting the margin to a specific value, as the inequality (4.12) does, and minimising the norm of $\boldsymbol{W}$ under this constraint.

$$\boldsymbol{W}.\boldsymbol{X} + b = 0 \tag{4.10}$$

$$y_i(\boldsymbol{W}.\boldsymbol{X}_i + b) > 0 \quad for\ \forall\ i = 1, \cdots, n \tag{4.11}$$

$$y_i(\boldsymbol{W}.\boldsymbol{X}_i + b) \geq 1 \quad for\ \forall\ i = 1, \cdots, n \tag{4.12}$$

However, as stated in section 4.7.1, medical datasets can rarely achieve the clean separation of the two classes needed in SVMs with hard margins; even if they do, using hard margins may well be risking overfitting. To overcome this issue SVMs with soft margins were developed, the crucial difference being that

they allowed data points to break inequality (4.12), meaning they could be on the wrong side of its supporting hyperplane (166). This was done by including a slack variable , ε_i, for each data point, as shown in inequality (4.13), which measures how far the data point is from satisfying inequality (4.12). As stated in section 4.7.1, the slack variables of potential hyperplanes are used to penalise them in order to discourage non-zero slack variables occurring and them being large when they do occur. This is done by adding them, scaled by the penalising parameter C, to the expression to be minimised. This gives (4.14), the expression to be minimised in the soft margin case. Notice $\|\boldsymbol{W}\|$, the norm of $\boldsymbol{W}$, has been replaced with $\frac{1}{2}\|\boldsymbol{W}\|^2$ as they are equivalent in the linear case and this overcomes the issue of $\|\boldsymbol{W}\|$ containing a square root.

$$y_i(\boldsymbol{W}.\boldsymbol{X}_i + b) \geq 1 - \varepsilon_i \quad for \ \forall \ i = 1, \cdots, n \quad , \varepsilon_i \geq 0 \tag{4.13}$$

$$\frac{1}{2}\|\boldsymbol{W}\|^2 + C\sum_{i=1}^{n}\varepsilon_i \tag{4.14}$$

The minimum of (4.14) subject to the constraint (4.13) can be solved by introducing Lagrange multipliers, α_i, and finding the stationary point of the Lagrange function, this is detailed elsewhere (168). Do note however, that calculating the solution to this minimisation problem involves substituting $\boldsymbol{W} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{X}_i$, where $\alpha_i = 0$ when the ith data point is not a support vector; this leads to $\|\boldsymbol{W}\|^2$ being replaced with $\sum_{i,j}\alpha_i\alpha_j y_i y_j(\boldsymbol{X}_i.\boldsymbol{X}_j)$.

The linear SVM included in this chapter was built in Rstudio using the *svm* function from the *e1071* library (169), with type set to 'C-classification' and kernel set to 'linear'. This minimises (4.14) subject to (4.13) using Platt's sequential minimal optimisation algorithm (170). The iterative process of minimising (4.14) stops once the reduction seen is less than the value chosen for the tolerance of termination criterion. The parameters of tolerance of termination criterion and C were varied using the default values as well as checking the performance of values above and below the default. Values of 0.1, 0.01, 0.001 and 0.0001 were assessed for the tolerance of termination criterion. These have been chosen as they span the default value, 0.001. C, the penalising parameter for non-zero slack values was evaluated at 0.1, 1, 10 and

100. The default value for C is 1. Finally, the class weights option of *svm* was utilised as there are roughly four times the number of non-cases compared to cases, which unadjusted for may lead to poor sensitivity which is of particular interest in this context. A weighting of 0.196 for non-cases and 0.804 for cases was used as this will lead to a balance in total weighting of cases and non-cases if the number of support vectors are even from each class.

4.7.2.2 Details of radial SVM

Non-linear SVMs can be built by utilising kernel functions to learn about the data in a feature space with a higher dimension than the number of explanatory variables, meaning non-linear relationships can be taken into account. This is achieved by calculating the inner products, the generalisation of dot products, of pairs of data points in that feature space using the kernel function, $K(\boldsymbol{X_i}, \boldsymbol{X_j})$ (167,168). The kernel function of vectors replaces the dot products used in the equations in section 4.7.2.1. The radial kernel function is given in (4.15).

$$K(\boldsymbol{X_i}, \boldsymbol{X_j}) = \exp(-\gamma \left\| \boldsymbol{X_i} - \boldsymbol{X_j} \right\|^2) \quad , \gamma > 0 \tag{4.15}$$

The radial SVM included in this chapter was built in Rstudio using the *svm* function from the *e1071* library (169), with type set to 'C-classification' and kernel set to 'radial'. The parameters of tolerance of termination criterion, C and γ were varied using the default values as well as checking the performance of values above and below the default. Values of 0.1, 0.01, 0.001 and 0.0001 were assessed for the tolerance of termination criterion. These have been chosen as they span the default value, 0.001. C, the penalising parameter for non-zero slack values was evaluated at 0.1, 1, 10 and 100. The default value for C is 1. The default for γ is, one over the number of explanatory variables, or 0.0667 in this case; this was assessed along with 0.25 and 0.01. As with the linear SVM a weighting of 0.196 for non-cases and 0.804 for cases was used. Due to the computational intensity of the radial SVM method, only the first of the 20 imputations was used in the development of the radial SVMs.

4.7.3 Results of SVM

As can be seen from Table 4.12 the linear SVM had good discrimination in both the internal and external datasets. The radial SVM had good discrimination for the internal dataset but this dropped considerably for the external dataset.

Table 4.12 AUROC of linear and radial SVMs on internal and external datasets with various penalty parameters (with γ and tolerance of termination criterion set to their default values)

Kernel	**Penalising parameter, C**	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier Score** (Outcome index variance)	**External Brier Score** (Outcome index variance)
Linear	0.1	0.701 (0.675, 0.727)	0.713 (0.691, 0.734)	0.178 (0.158)	0.155 (0.158)
	1	0.701 (0.675, 0.727)	0.713 (0.691, 0.734)	0.178 (0.158)	0.154 (0.158)
	10	0.701 (0.675, 0.727)	0.713 (0.691, 0.734)	0.178 (0.158)	0.154 (0.158)
	100	0.701 (0.675, 0.727)	0.713 (0.691, 0.734)	0.178 (0.158)	0.154 (0.158)
Radial	0.1	0.687 (0.659, 0.715)	0.628 (0.603, 0.6528)	0.179 (0.158)	0.162 (0.158)
	1	0.687 (0.659, 0.715)	0.628 (0.603, 0.6528)	0.179 (0.158)	0.162 (0.158)
	10	0.687 (0.659, 0.715)	0.628 (0.603, 0.6528)	0.179 (0.158)	0.162 (0.158)
	100	0.687 (0.659, 0.715)	0.628 (0.603, 0.6528)	0.179 (0.158)	0.162 (0.158)

The results from the comparison here support the use of the linear SVM method but not the radial SVM method due to its lack of external validity.

To summarise, of all the methods compared using the multiply imputed internal dataset, logistic regression, boosted decision tree and linear SVM were the only methods which achieved acceptable levels of discrimination, with AUROCs greater than 0.70 in both the internal and external datasets. However, the boosted decision tree method resulted in models with poor calibration. The findings of this main analysis are discussed in detail in section 4.10, alongside

the findings of the sensitivity analysis, with both statistical performance and practical issues being considered.

4.8 Sensitivity analysis

4.8.1 Candidate variables

The sensitivity analysis included in this section was carried out to ensure none of the methods or their extensions were disadvantaged due to being built on a multiply imputed dataset. The sensitivity analysis only used data of individuals who have values for the all of the following variables:

- Outcome (normal blood glucose/NDH or undiagnosed T2DM by oral glucose tolerance test (OGTT))
- Age (years)
- BMI (Kg/m^2)
- Ethnicity (White European/Other ethnicity)
- Sex (Male/Female)
- Waist circumference (cm)
- 1^{st} degree family history of diabetes (yes/no)
- History of hypertension or antihypertensive use (yes/no)

95.4% of individuals aged 40-75 years old in the ADDITION-Leicester dataset had the data required to be included in this sensitivity analysis, 6,099 individuals. The same techniques as were used in the main analysis were implemented here, with the same parameters being used, apart from when the recommended value of a parameter is based on the number of candidate variables, which reduced to seven for this analysis. The decrease in the number of candidate variables resulted in the number of individuals included in the external dataset, STAR, increasing to 3,173.

4.8.2 Logistic regression

Performing stepwise backward elimination on the seven candidate variables included in this sensitivity analysis resulted in the logistic regression model detailed in Table 4.13. All variables included in this model had the same relationship with the outcome as reported in the literature, and thus were chosen to stay in the rest of the models built in this section. Some literature suggests males are at an increased risk of the outcome (171,172); however building another model with this variable included along with the rest resulted in

a non-significant increase in risk for females. For this reason sex was kept out of the rest of the logistic regression models built in this analysis.

Table 4.13 Logistic regression model selected by automatic backwards elimination with continuous variable kept continuous (sensitivity analysis)

Variable		Coefficient	95% CI	P Value
Age (years)		0.0474	0.039, 0.055	<0.001
Ethnicity	White	Reference group		
	Other	0.794	0.63, 0.96	<0.001
First degree family	No	Reference group		
history of T2DM	Yes	0.362	0.20, 0.53	<0.001
Waist circumference		0.0227	0.014, 0.031	<0.001
(cm)				
Hypertension	No	Reference group		
	Yes	0.405	0.25, 0.56	<0.001
BMI (kg/m^2)		0.0407	0.020, 0.062	<0.001

Considering interactions and quadratic terms lead to the interaction of hypertension and family history of diabetes being included along with a quadratic term for BMI. Table 4.14 shows having both a family history of diabetes and hypertension increases the chances of having NDH or undiagnosed T2DM compared to having neither or one of these risk factors alone.

Table 4.14 Logistic regression model with continuous variables kept continuous and interactions and quadratic terms considered (sensitivity analysis)

Variable		Coefficient	95% CI	P Value
Age (years)		0.0469	0.039, 0.055	<0.001
BMI (kg/m^2)		0.153	0.048, 0.26	0.004
Squared BMI		-0.00173	-0.0033, -0.00016	0.031
Ethnicity	White	Reference group		
	Other	0.783	0.62, 0.95	<0.001
First degree family	No	Reference group		
history of T2DM	Yes	0.162	-0.045, 0.37	0.124
Hypertension	No	Reference group		
	Yes	0.262	0.087, 0.44	0.003
Waist circumference (cm)		0.0215	0.013, 0.030	<0.001
Interaction: Hypertension & Family History		0.549	0.21,0.89	0.001

Table 4.15 displays the model and associated risk score when the continuous variables included are grouped, as can be seen all groups for each variable are significant.

Table 4.15 Logistic regression model and scoring system for risk assessment tool with grouped continuous variables (sensitivity analysis)

Variable	Grouping	Coefficient	95% CI	P Value	Scoring
Age (years)	40-49	Reference group			0
	50-59	0.412	0.21, 0.62	<0.001	4
	60-69	0.854	0.65, 1.06	<0.001	9
	70+	1.16	0.92, 1.41	<0.001	12
Ethnicity	White	Reference group			0
	Other	0.746	0.59, 0.91	<0.001	7
First degree family history of T2DM	No	Reference group			0
	Yes	0.338	0.17, 0.50	<0.001	7
Waist circumference (cm)	<90	Reference group			0
	90-99	0.483	0.28, 0.68	<0.001	5
	100-109	0.600	0.36, 0.84	<0.001	6
	>109	0.871	0.57, 1.17	<0.001	9
BMI (kg/m^2)	<25	Reference group			0
	25-29	0.237	0.023, 0.45	0.030	2
	30-34	0.409	0.146, 0.67	0.002	4
	≥35	0.707	0.380, 1.03	<0.001	7
Hypertension	No	Reference group			0
	Yes	0.430	0.28, 0.58	<0.001	4

Table 4.16 shows the discrimination and calibration of the RATs built using logistic regression in this sensitivity analysis were similar to the main analysis. Good levels of discrimination were seen in both the internal and external datasets, though calibration was a little higher in the external data.

Table 4.16 Discrimination and calibration of logistic regression risk assessment tools developed in the sensitivity analysis and main analysis

Development data	**Risk assessment tool**	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
Complete-case data (seven candidate variables and outcome)	Appropriate for electronic platform*	0.697 (0.641, 0.753)	0.706 (0.684, 0.728)	0.135 (0.145)	0.148 (0.161)
	Continuous variables grouped	0.686 (0.631, 0.740)	0.698 (0.676, 0.720)	0.136 (0.145)	0.148 (0.161)
Multiple imputed data (15 candidate variables)	Appropriate for electronic platform	0.701 (0.655, 0.748)	0.714 (0.692, 0.736)	0.145 (0.158)	0.144 (0.158)
	Continuous variables grouped	0.723 (0.685, 0.762)	0.702 (0.680, 0.724)	0.143 (0.158)	0.145 (0.158)

*(interaction of family history with hypertension and squared term for BMI)

4.8.3 Basic decision tree and extensions

The majority of the results are very consistent with the main analysis. The basic decision trees and boosted decision trees fitted have similar results to the main analysis. Although, the difference between the internal and external AUROCs for the basic decision trees was a little larger in the complete-case analysis.

Table 4.17 Internal and external AUROCs and Brier scores of decision trees with various weighting for cases-to-non-cases in the sensitivity analysis and main analysis

Development data	**Weighting of cases compared to non-cases**	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
Complete-case data (seven candidate variables and outcome)	**1 :1**	0.500 (0.500, 0.500)	0.500 (0.500, 0.500)	0.145 (0.145)	0.161 (0.161)
	2.5 :1	0.625 (0.568, 0.682)	0.615 (0.592, 0.637)	0.167 (0.145)	0.177 (0.161)
	5 :1	0.649 (0.599, 0.699)	0.626 (0.602, 0.649)	0.240 (0.145)	0.252 (0.161)
	7.5 :1	0.599 (0.559, 0.639)	0.583 (0.566, 0.601)	0.319 (0.145)	0.319 (0.161)
	10 :1	0.598 (0.560, 0.636)	0.588 (0.571, 0.604)	0.375 (0.145)	0.368 (0.161)
Multiple imputed data (15 candidate variables)	**1 :1**	0.500 (0.500, 0.500)	0.500 (0.500, 0.500)	0.158 (0.158)	0.158 (0.158)
	2.5 :1	0.630 (0.570, 0.690)	0.627 (0.603, 0.652)	0.182 (0.158)	0.187 (0.158)
	5 :1	0.636 (0.580, 0.692)	0.637 (0.613, 0.660)	0.262 (0.158)	0.275 (0.158)
	7.5 :1	0.592 (0.555, 0.628)	0.582 (0.565, 0.600)	0.341 (0.158)	0.346 (0.158)
	10 :1	0.589 (0.552, 0.626)	0.584 (0.567, 0.602)	0.395 (0.158)	0.399 (0.158)

Decision trees with 10 cross-validations, minimum number of observations in terminal node=10 and complexity parameter=0.01

The reduced size of the development dataset for the sensitivity analysis allowed the number of trees to be increased to 20,000 in one model developed. The AUROC decreased slightly in both the internal and external data for this further increase in number of trees, showing that the number of trees needed for optimum discrimination had also been reached for the complete-case development data.

Table 4.18 Internal and external AUROC and Brier scores for boosted decision trees with varying number of iterations in the sensitivity analysis and main analysis

Development Data	**Number of Trees**	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
Complete-case data (seven candidate variables and outcome)	**500**	0.665 (0.642, 0.688)	0.663 (0.639, 0.686)	0.250 (0.145)	0.257 (0.161)
	1,000	0.673 (0.650, 0.696)	0.674 (0.651, 0.697)	0.243 (0.145)	0.244 (0.161)
	5,000	0.694 (0.671, 0.717)	0.696 (0.674, 0.718)	0.228 (0.145)	0.237 (0.161)
	10,000	0.695 (0.671, 0.718)	0.694 (0.672, 0.716)	0.225 (0.145)	0.239 (0.161)
	20,000	0.691 (0.667, 0.715)	0.688 (0.665, 0.710)	0.223 (0.145)	0.240 (0.161)
Multiple imputed data (15 candidate variables)	**500**	0.659 (0.633, 0.685)	0.664 (0.640, 0.687)	0.271 (0.158)	0.274 (0.158)
	1,000	0.670 (0.644, 0.696)	0.678 (0.655, 0.701)	0.265 (0.158)	0.269 (0.158)
	5,000	0.700 (0.674, 0.723)	0.703 (0.681, 0.725)	0.247 (0.158)	0.254 (0.158)
	10,000	0.703 (0.677, 0.729)	0.703 (0.681, 0.725)	0.243 (0.158)	0.250 (0.158)

Boosted decision tree with 5:1 case-to-control weighting, shrinkage=0.001, fraction=0.5, maximum depth=2, minimum number of observations in terminal node=10, cross-validations=10

The bagged decision tree method produced very similar AUROCs and Brier scores in the internal and external datasets to the main analysis, with a maximum depth of five, once again allowing the best AUROC to be achieved.

Table 4.19 Internal and external AUROC and Brier scores for bagged decision trees with varying maximum depths in the sensitivity analysis and main analysis

Development data	**Maximum depth of trees**	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
Complete-case data (seven candidate variables and outcome)	2	0.649 (0.621, 0.677)	0.657 (0.634, 0.680)	0.151 (0.145)	0.154 (0.161)
	5	0.680 (0.652, 0.707)	0.678 (0.656, 0.701)	0.148 (0.145)	0.151 (0.161)
	7	0.678 (0.651, 0.706)	0.675 (0.653, 0.698)	0.149 (0.145)	0.152 (0.161)
	10	0.677 (0.645, 0.704)	0.673 (0.651, 0.696)	0.149 (0.145)	0.153 (0.161)
Multiple imputed data (15 candidate variables)	2	0.645 (0.621, 0.677)	0.659 (0.635, 0.683)	0.152 (0.158)	0.151 (0.158)
	5	0.680 (0.652, 0.707)	0.682 (0.659, 0.706)	0.148 (0.158)	0.148 (0.158)
	7	0.678 (0.651, 0.706)	0.681 (0.658, 0.704)	0.149 (0.158)	0.149 (0.158)
	10	0.677 (0.649, 0.704)	0.680 (0.657, 0.703)	0.149 (0.158)	0.149 (0.158)

Bagged decision trees with 5:1 case-to-control weighting, complexity parameter=0.001, bootstrapped samples=100, minimum node size=10

The number of variables considered at each stage of the random forest was set to three in this sensitivity analysis, as the number of candidate variables was reduced. The random forest method had lower Brier scores in the sensitivity analysis compared to the main analysis. The discrimination was again an issue, with 0.66 being the best AUROC achieved in the external dataset.

Table 4.20 Internal and external AUROCs and Brier scores for random forests with varying minimum number of observations in terminal nodes in the sensitivity analysis and main analysis

Development data	**Minimum size of terminal node**	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier score** (Outcome index variance)	**External Brier score** (Outcome index variance)
Complete-case data (seven candidate variables and outcome) [a]	**1**	0.667 (0.690, 0.643)	0.638 (0.614, 0.661)	0.139 (0.145)	0.165 (0.161)
	10	0.674 (0.651, 0.698)	0.649 (0.625, 0.672)	0.170 (0.145)	0.225 (0.161)
	100	0.668 (0.645, 0.691)	0.659 (0.636, 0.682)	0.580 (0.145)	0.567 (0.161)
Multiple imputed data (15 candidate variables) [b]	**1**	0.635 (0.612, 0.658)	0.655 (0.631, 0.679)	0.284 (0.158)	0.221 (0.158)
	10	0.655 (0.631, 0.678)	0.680 (0.656, 0.703)	0.446 (0.158)	0.370 (0.158)
	100	0.638 (0.614, 0.661)	0.678 (0.655, 0.701)	0.638 (0.158)	0.657 (0.158)

[a]Random Forests with 5:1 case-to-control weighting, three variables considered at each split, trees=500

[b]Random forests with 5:1 case-to-control weighting, four variables considered at each split, trees=500

4.8.4 Support Machine Vector

Table 4.21 shows the linear SVM method produced acceptable levels of discrimination, with the AUROC being a little lower in the external dataset compared to internal dataset. On the other hand, the radial SVM models produced AUROCs of 0.66 in the external dataset. These results are consistent with the findings of the main analysis.

Table 4.21 AUROC of linear and radial SVMs on internal and external datasets with various penalty parameters (with γ and tolerance of termination criterion set to their default values) (sensitivity analysis)

Kernel	**Penalising parameter, C**	**Internal cross-validated AUROC** (95% CI)	**External AUROC** (95% CI)	**Internal cross-validated Brier Score** (Outcome index variance)	**External Brier Score** (Outcome index variance)
Linear	0.1	0.695 (0.670, 0.719)	0.683 (0.666, 0.700)	0.172 (0.145)	0.156 (0.161)
	1	0.695 (0.670, 0.719)	0.683 (0.666, 0.700)	0.172 (0.145)	0.156 (0.161)
	10	0.695 (0.670, 0.719)	0.683 (0.666, 0.700)	0.172 (0.145)	0.156 (0.161)
	100	0.695 (0.670, 0.719)	0.683 (0.666, 0.700)	0.172 (0.145)	0.156 (0.161)
Radial	0.1	0.684 (0.658, 0.711)	0.657 (0.634, 0.681)	0.173 (0.145)	0.158 (0.161)
	1	0.684 (0.658, 0.711)	0.657 (0.634, 0.681)	0.173 (0.145)	0.158 (0.161)
	10	0.684 (0.658, 0.711)	0.657 (0.634, 0.681)	0.173 (0.145)	0.158 (0.161)
	100	0.684 (0.658, 0.711)	0.657 (0.634, 0.681)	0.173 (0.145)	0.158 (0.161)

4.9 Resampling study on the effects of the sample size of the development dataset on performance of methods

The sample size, or more specifically the number of events per candidate variable (EPV), is known to be an important element of developing a stable RAT (131). RATs which use too few EPV are prone to overfitting on the development dataset and overestimating predictive performance (173,174). Previous resampling studies considering the effects of the number of EPV on cross-sectional prediction models have tended to concentrate on the accuracy of the parameter estimates in logistic regression models (173,175-178). This section details a resampling study focusing on the effect of the number of EPV on the predictive performance of RATs developed by each of the methods considered in the empirical comparison in this chapter.

4.9.1 Methods

Models were developed using the ADDITION-Leicester dataset, restricted to cases with complete data for the outcome and the seven predictors included in the LSA score and externally validated using the STAR dataset, these two datasets have been detailed earlier in section 4.2.2. Complete case analysis has been used for computational reasons to allow the impact of sample size on the relative performance of the different methods, some of which are computationally intensive, to be assessed; consistent results were found in the earlier empirical comparison of methods in both the multiply imputed and complete case datasets.

For each sample size considered, 1,000 samples were drawn with replacement from the ADDITION-Leicester dataset. The event rate was fixed at 17.6% for every sample by stratified sampling according to the outcome. Six sample sizes were considered corresponding to the following numbers of EPV: 2, 5, 10, 15, 20 and 50.

One prediction model for each of the methods included in the empirical comparison in this chapter was assessed for each sample size. The models compared were:

- Logistic regression: Backward elimination starting with the full model with a p-value of 0.05 or less required to stay in the model.

- Decision tree: 5:1 case-to-control weighting, 10 as the minimum number of observations in terminal node, cross-validations=10 and complexity parameter=0.01
- Boosted decision tree: 5,000 trees, 5:1 case-to-control weighting, shrinkage=0.001, fraction=0.5, maximum depth=2, minimum number of observations in terminal node=10, cross-validations=10
- Bagged decision trees: maximum depth=5, 5:1 case-to-control weighting, complexity parameter=0.001, bootstrapped samples=100, minimum node size=10
- Random forest: minimum size of terminal node=10, 5:1 case-to-control weighting, four variables considered at each split, trees=500
- Linear SVM: penalising parameter=1, γ and tolerance of termination criterion set to their default values
- Radial SVM: penalising parameter=1, γ and tolerance of termination criterion set to their default values

These were the models which performed the best out of those applied in the empirical comparison for each method, with the exception of the logistic regression model specified above. This is because the techniques considered for logistic regression in the empirical comparison involved a human element, meaning that repeating the required number of times to allow for their inclusion in a resampling study was not feasible. The automatic logistic regression method was carried in the R package *rms* using the *fastbw* command (179). The R packages used for the other methods have been detailed in the empirical comparison, earlier in this chapter.

For each model developed using each sample the AUROC and Brier score were calculated for the internal sample using stratified 10-fold cross-validation and for the full external dataset. This allowed the metrics, yielded both internally and externally for each method, to be compared as the number of EPV varied. The mean percentage bias and Root Mean Square Error (RMSE) of the AUROCs and Brier scores across the 1,000 samples compared to the AUROC and Brier score achieved when using the full internal dataset to develop each model were calculated for each method and sample size. Equations (4.16) and (4.17) give the definition of the mean percent bias and RMSE respectively,

where S_i is the value of the metric in the i[th] sample and F is the metric yielded when using the whole internal dataset to develop the model.

$$Mean\ percent\ bias = \frac{100}{n}\sum_{i=1}^{n}\frac{S_i - F}{F} \tag{4.16}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - F)^2} \tag{4.17}$$

To assess the stability of the discrimination, the proportion of external AUROCs within 1%, 2.5%, 5%, and 10% of their corresponding cross-validated internal AUROC was calculated for each method and sample size considered.

4.9.2 Results

Figure 4.7 shows that the internal AUROCs yielded from logistic regression models with small numbers of EPV varied greatly with the simulations with two EPV producing AUROCs ranging between 0.450 and 0.930. As expected, the range of the AUROCs reduced as the number of EPV is increased. The external AUROCs followed the same pattern, however reassuringly the range of the AUROCs did not extend above 0.71; the external AUROC achieved using the whole internal dataset to develop this model. Instead the AUROCs produced converged to this value as the number of EPV increased.

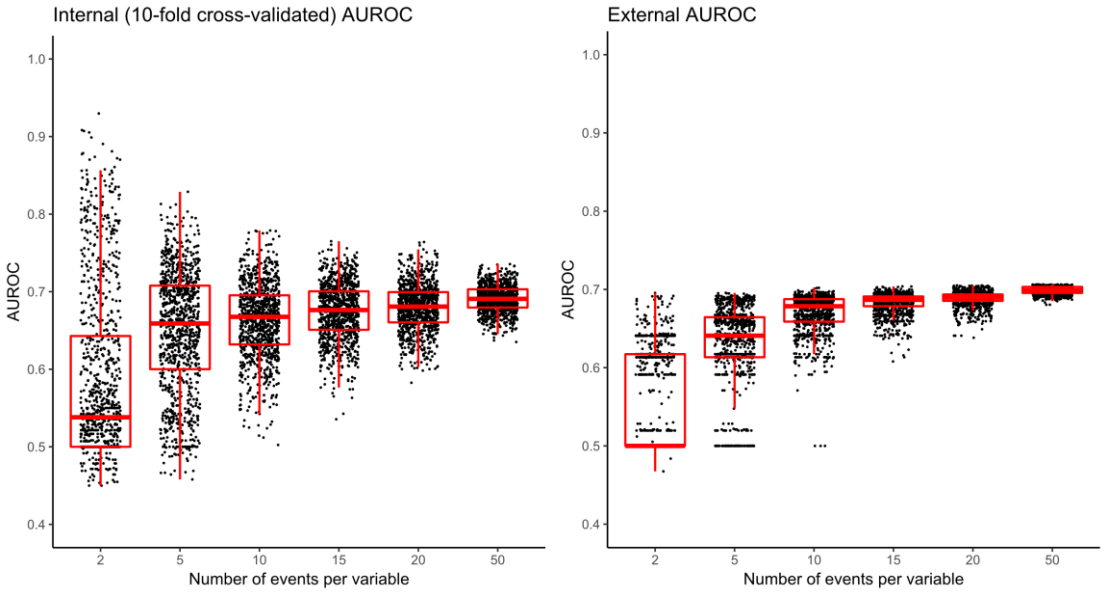


Figure 4.7 AUROCs of logistic regression RATs developed using samples with different numbers of EPV

Figure 4.8 demonstrates that the Brier scores of the simulations tended towards the Brier scores returned from implementing this logistic regression method using the full development dataset, 0.136 and 0.148 in the internal and external dataset respectively, as the number of EPV used in development were increased. The range of the external Brier scores did not go below 0.148.

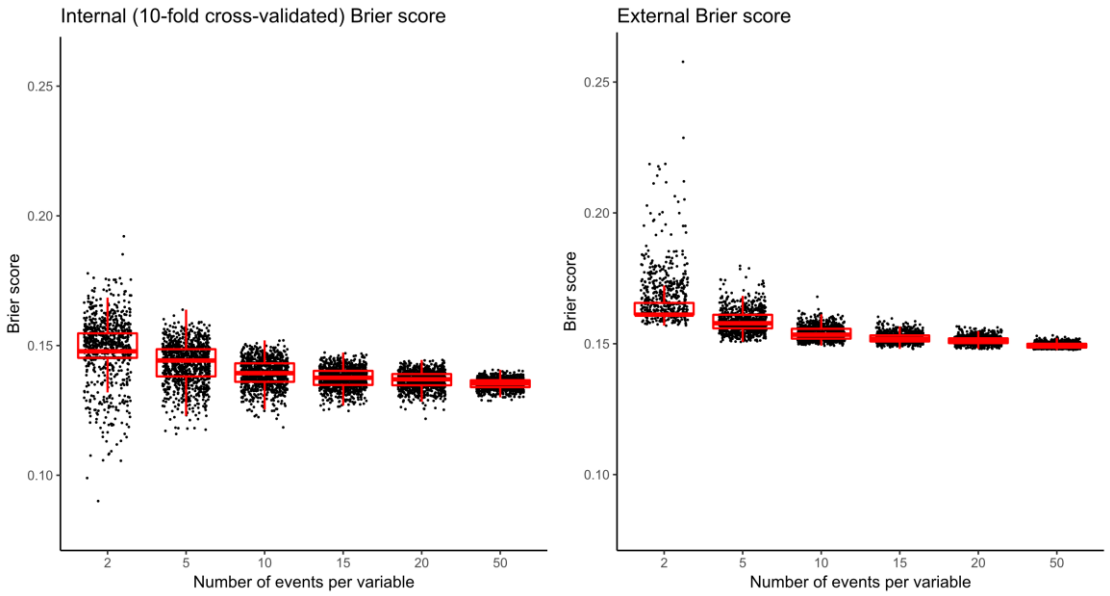


Figure 4.8 Brier scores of logistic regression RATs developed using samples with different numbers of EPV

Figure 4.9 shows that the range of the AUROCs decreased with increasing numbers of EPV. The majority of the external AUROCs were below 0.65. The medians for the different numbers of EPV increase towards 0.626, the external AUROC of the model development with the full internal dataset, as the number of EPV does. Figure 4.10 displays that the range of the Brier scores decreased with decreasing numbers of EPV.

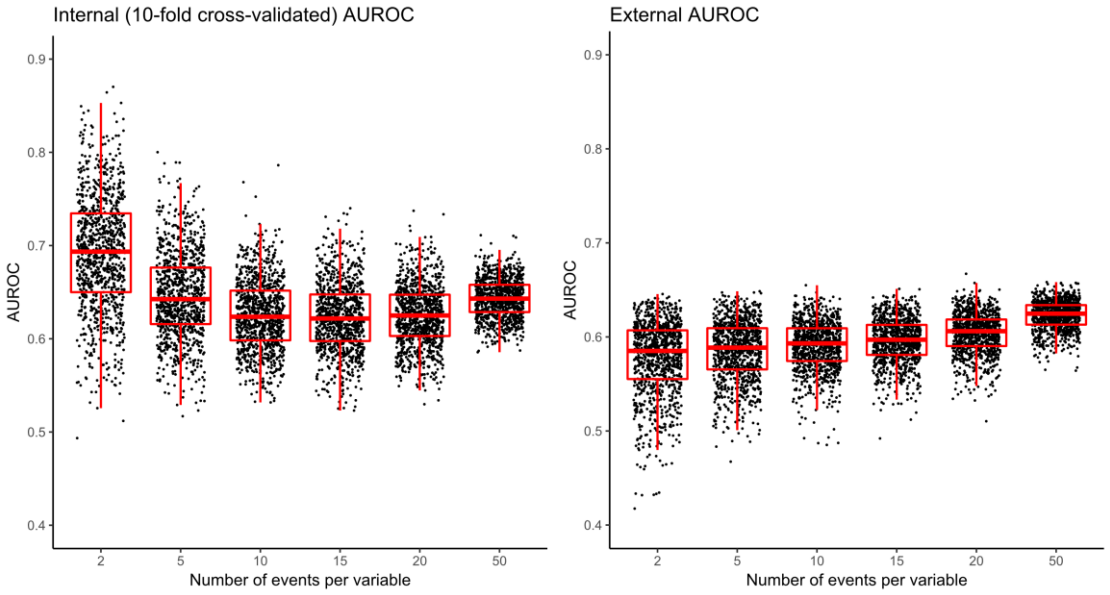


Figure 4.9 AUROCs of decision trees developed using samples with different numbers of EPV

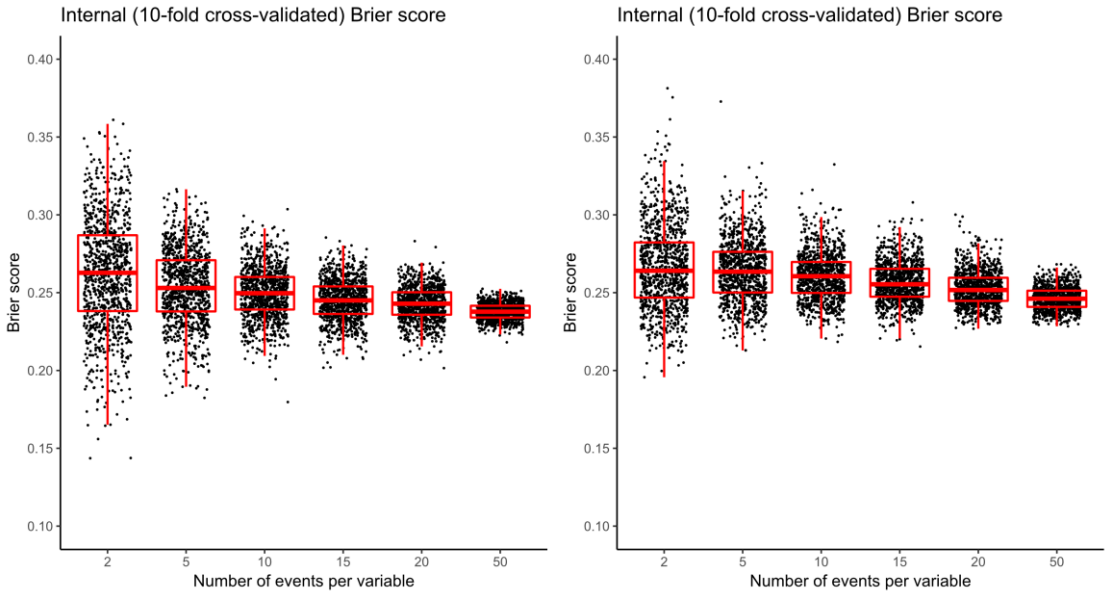


Figure 4.10 Brier scores of decision trees developed using samples with different numbers of EPV

Figure 4.11 indicates that the variability of the AUROCs produced by the boosted decision trees was reduced when the number of EPV was higher. The maximum external AUROC was 0.700, only marginally above the external AUROC of 0.696 produced when using the full internal dataset to develop this model. Figure 4.12 displays that the range of the Brier scores decreased with increasing numbers of EPV, surprisingly it showed that the average Brier score increased for higher numbers of EPV both internally and externally.

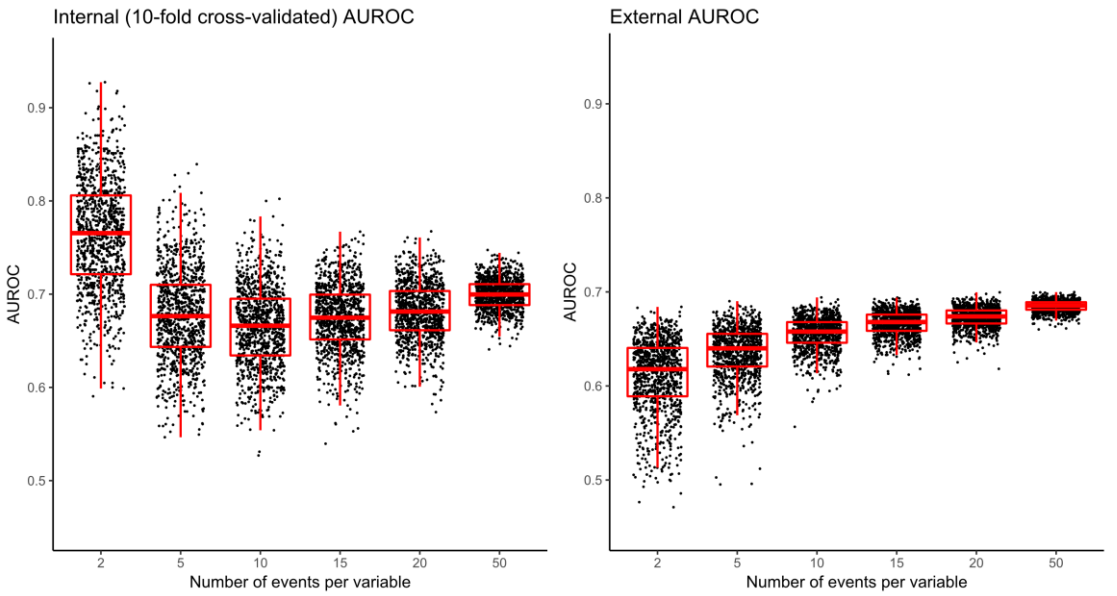


Figure 4.11 AUROCs of boosted decision trees developed using samples with different numbers of EPV

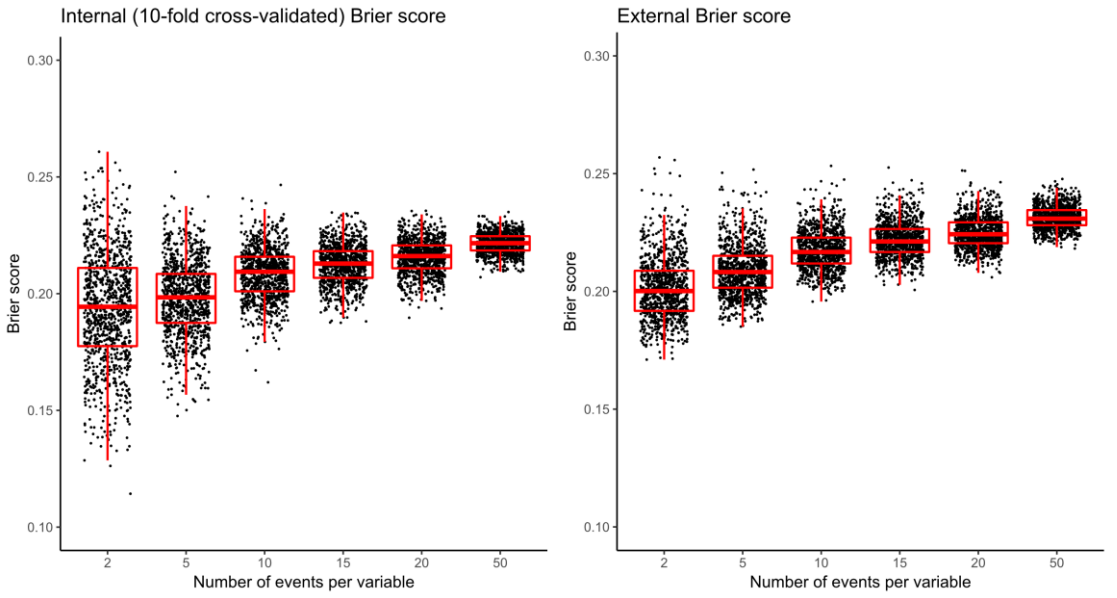


Figure 4.12 Brier scores of boosted decision trees developed using samples with different numbers of EPV

Figure 4.13 shows that both the internal and external AUROCs produced using the bagged decision tree method became more consistent as the number of EPV were increased. The majority of the external AUROCs were less than 0.678, the external AUROC of the model built using method with the full internal dataset, although the external AUROCs tended towards this value as the number of EPV are increased.

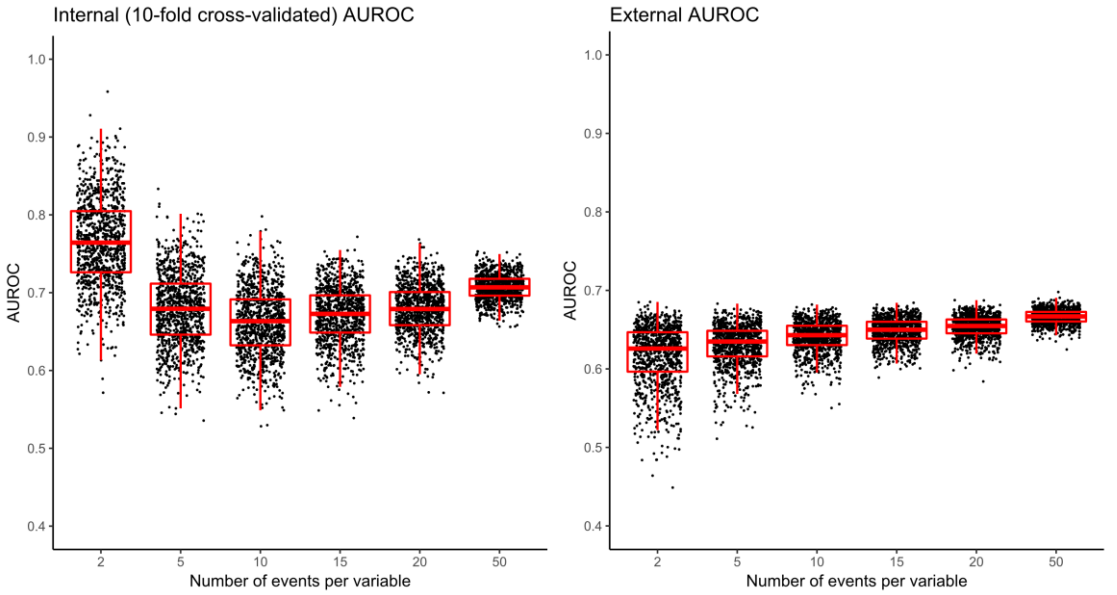


Figure 4.13 AUROCs of bagged decision trees developed using samples with different numbers of EPV

Figure 4.14 displays that as the number of EPV increased both the variability and averages of the Brier scores reduced. Few external Brier scores were below the 0.151 yielded from the model which was developed using the whole dataset.

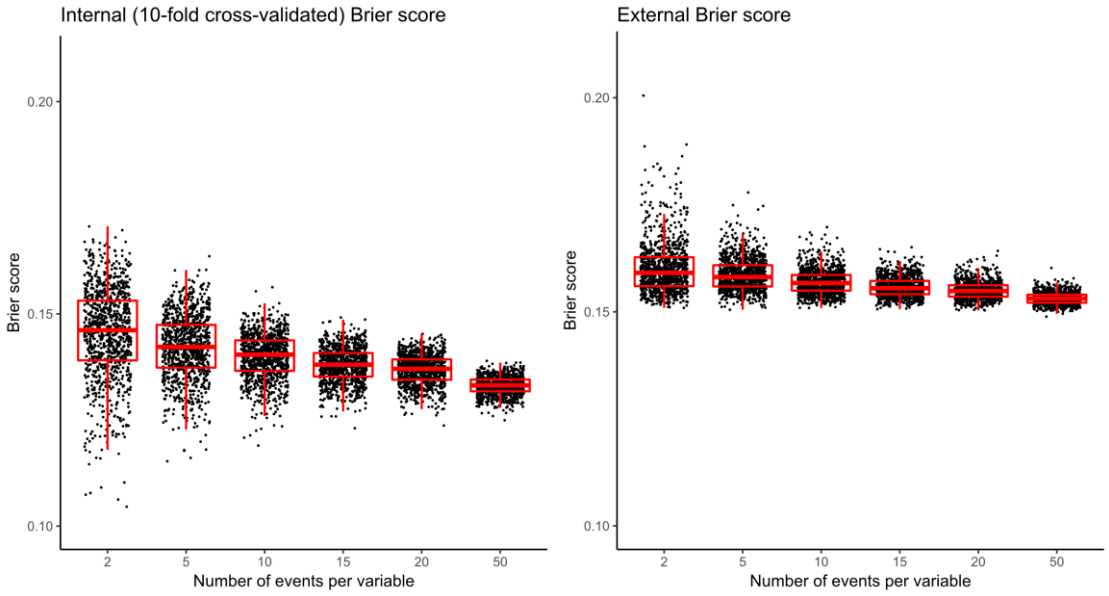


Figure 4.14 Brier scores of bagged decision trees developed using samples with different numbers of EPV

Figure 4.15 illustrates that the averages of the external AUROCs tended toward the external AUROC produced by the corresponding random forest model developed using the whole dataset, 0.649. However, the internal AUROCs produced when there were 50 EPV were all above the 0.674; the internal AUROC of the model developed using the whole dataset. Figure **4.15**e 4.16 displays that the variability of the Brier scores decreases with increasing numbers of EPV, however the majority of the internal scores were markedly above the 0.170 yielded from the model developed using the whole dataset.

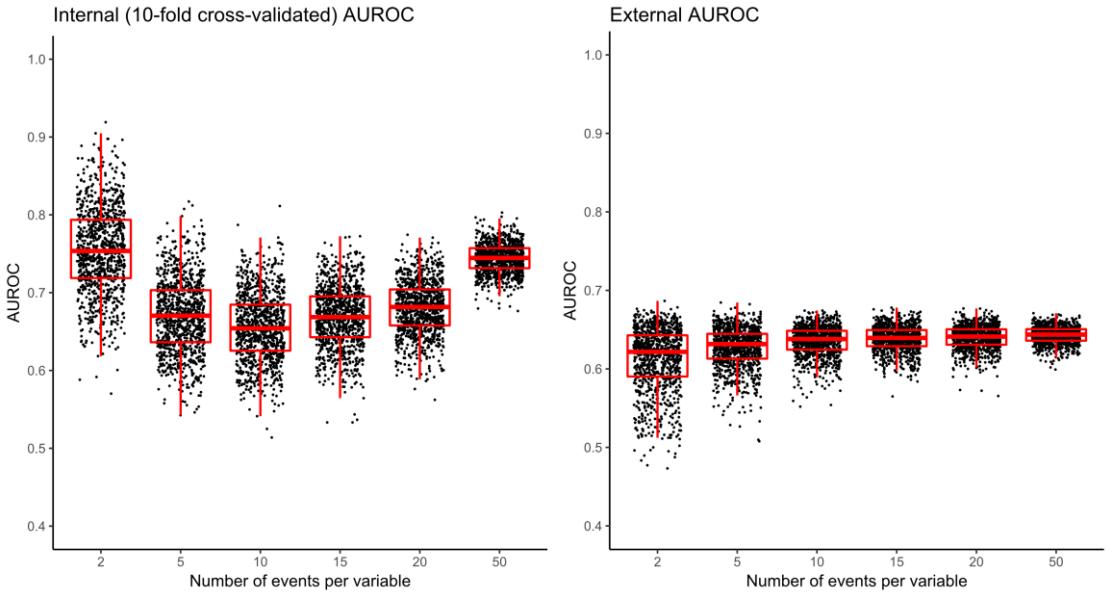


Figure 4.15 AUROCs of random forests developed using samples with different numbers of EPV

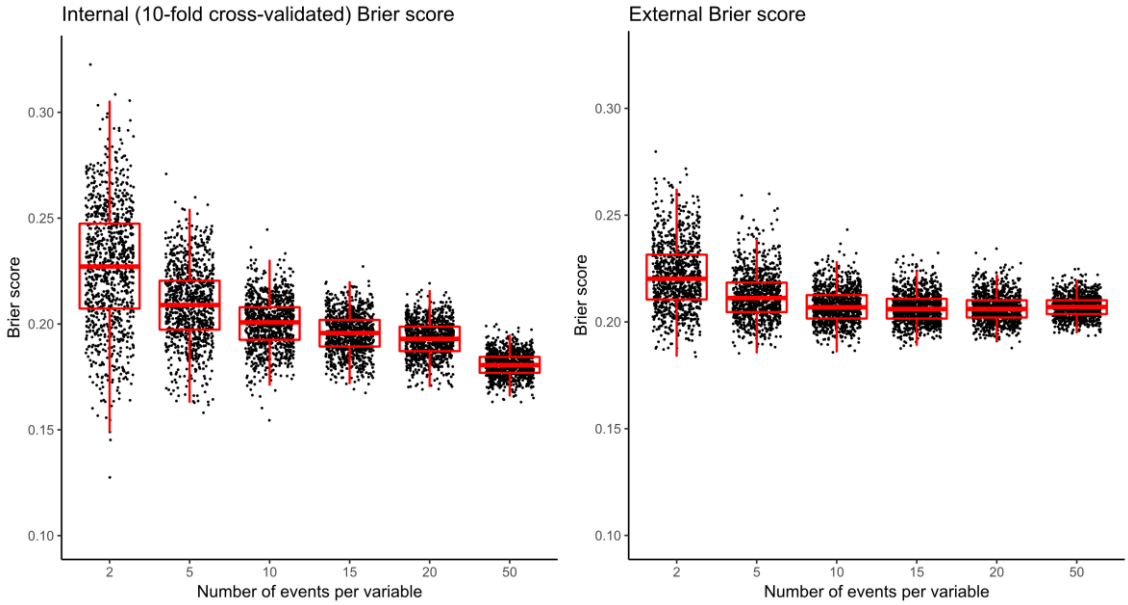


Figure 4.16 Brier scores of random forests developed using samples with different numbers of EPV

Figure 4.17 and Figure 4.18 show that the variability of both the AUROCs and Brier scores reduced as the number of EPV increased. The majority of the external AUROCs were less than 0.70, while most external Brier scores were above 0.148, indicating that models produced by the simulations which externally outperform the model developed using the whole internal dataset do so by a small margin. The model developed using the full dataset has an external AUROC and Brier score of 0.683 and 0.156 respectively.

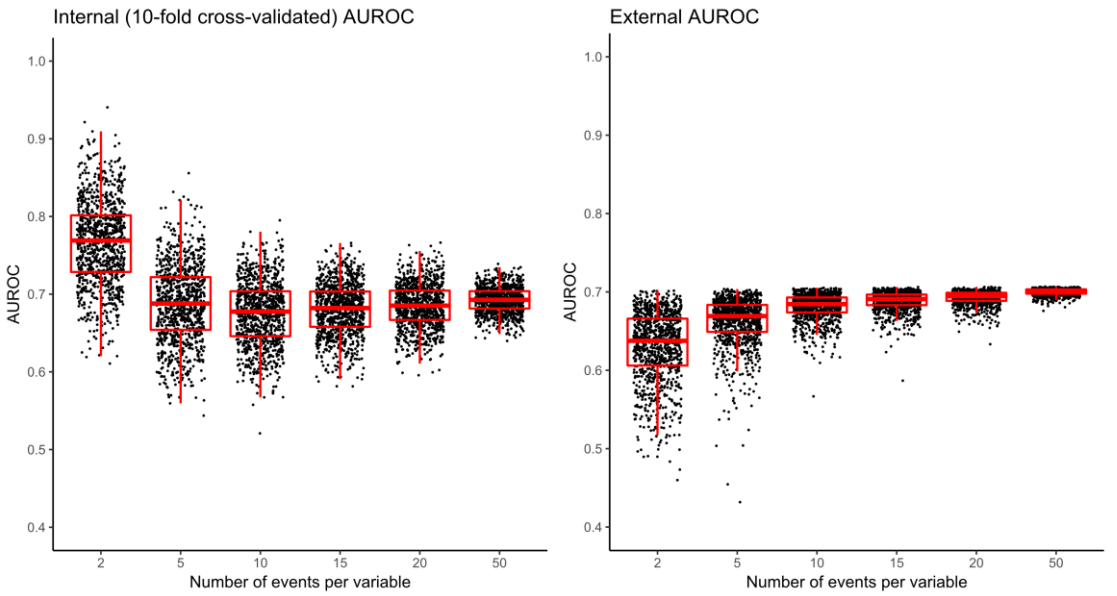


Figure 4.17 AUROCs of linear SVMs developed using samples with different numbers of EPV

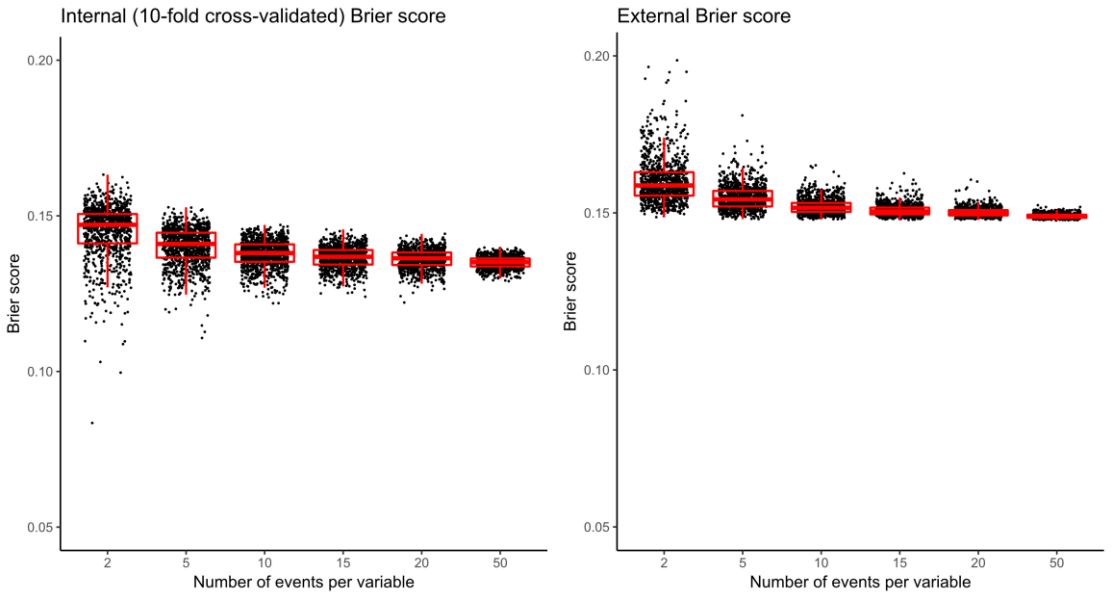


Figure 4.18 Brier scores of linear SVMs developed using samples with different numbers of EPV

Figure 4.19 and Figure 4.20 display that the variablity of the AUROCs and Brier scores decreased as the number of EPV was increased. There were numerous external AUROCs between 0.657, the value attained under the model developed with the full internal dataset, and 0.700. The vast majority of the internal Brier scores were less than the 0.173 observed for the corresponding radial SVM developed using the full internal dataset.

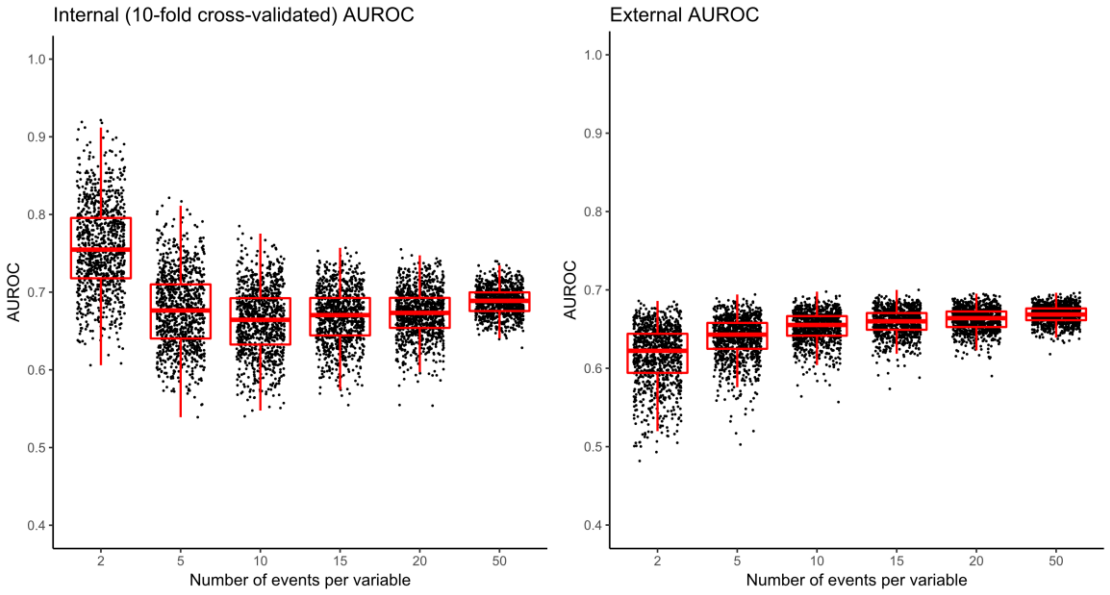


Figure 4.19 AUROCs of radial SVMs developed using samples with different numbers of EPV

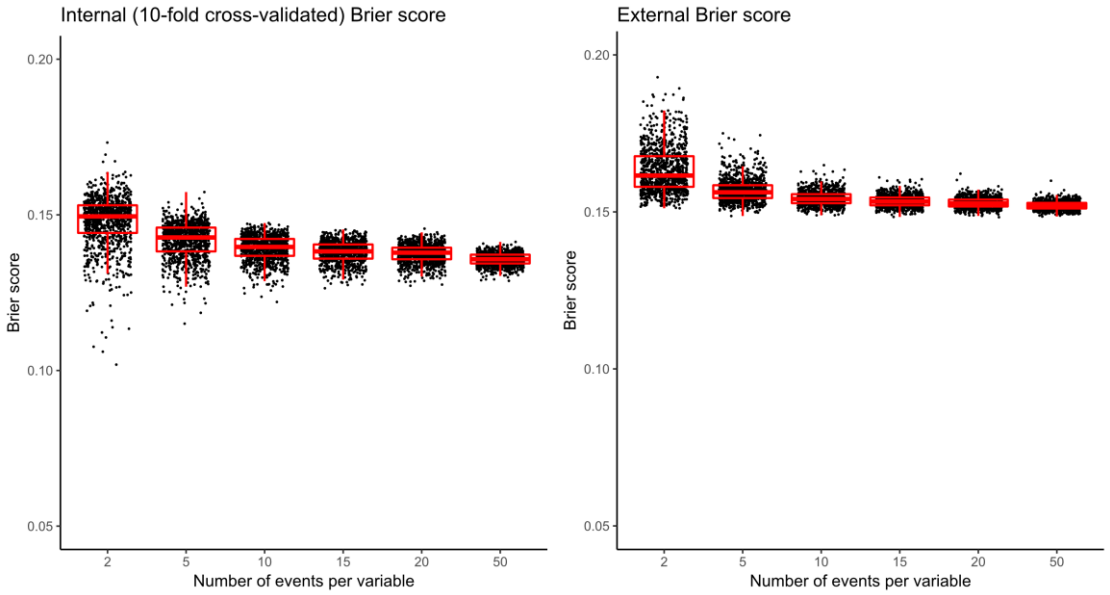


Figure 4.20 Brier scores of radial SVMs developed using samples with different numbers of EPV

Figure 4.21 indicates that for two EPV, all methods except the logistic regression had a large positive mean bias for internal AUROC; however all methods had a negative mean bias for internal AUROC once the number of EPV had been increased to 10, all mean biases being between -2.5% and -5.0% at this point. The internal mean biases tended to rise as the number of EPV increased from 10 to 50, with most methods having a mean bias between -1% and 1% for 50 EPV. Alarmingly, the random forest has a mean bias of 10.3% for 50 EPV, indicating unreliability of the random forest in the internal dataset.

Apart from the basic decision tree method, the mean biases of the external AUROCs all followed a similar pattern. The plot indicates that for small numbers of EPV they produced markedly worse external AUROCs than could be achieved by the method using the full internal dataset, however as the number of EPVs increased the mean bias increased towards zero.

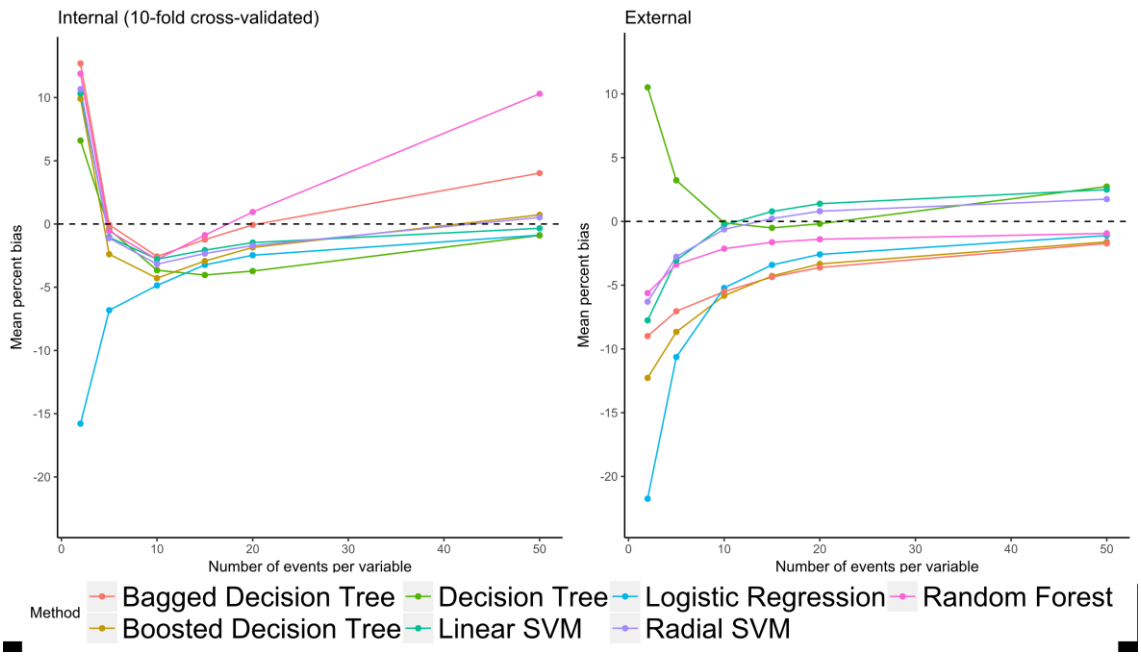


Figure 4.21 Mean percentage bias of AUROCs produced using various sample sizes compared to using full internal dataset to development models for each method

Figure 4.22 shows that the mean percent bias decreased with increasing numbers of EPV for all methods except the boosted decision tree, in which the opposite occurred. Noticeably the mean biases for the internal Brier scores of both the linear and radial SVM scores were extremely low, being around -20.0% once the numbers of EPV was at least five, indicating that the internal Brier score may have been overestimated in the empirical comparison. However, somewhat reassuringly the mean bias was only around -5.0% for these models in the external data, indicating the utility of an external validation for RATs built using these methods. As with the mean biases of the AUROCs, the logistic regression mean biases were very similar in the internal and external data.

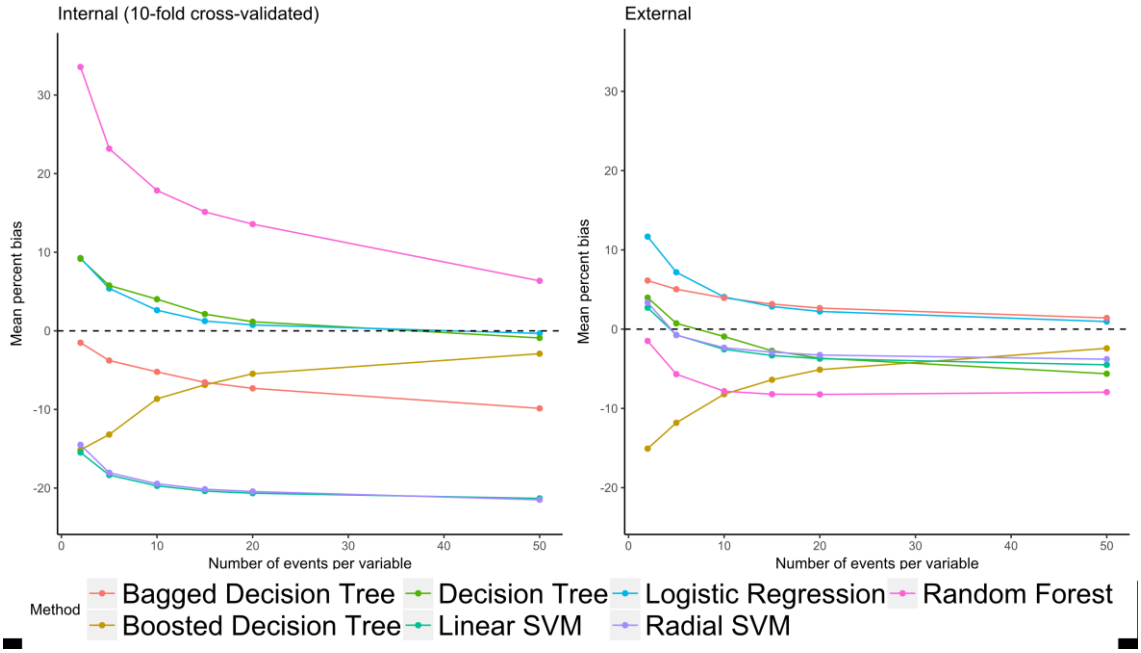


Figure 4.22 Mean percentage bias of Brier scores produced using various sample sizes compared to using full internal dataset to development models for each method

Figure 4.23 displays that the RMSE of the both the internal and external AUROCs decreased as the number of EPV increased for the majority of methods. Significantly, the random forest did not follow this pattern for the RMSE of the internal AUROC. Generally, the linear and radial SVM had the lowest RMSE. The plot shows that the logistic regression had the highest RMSEs for low numbers of EPV.

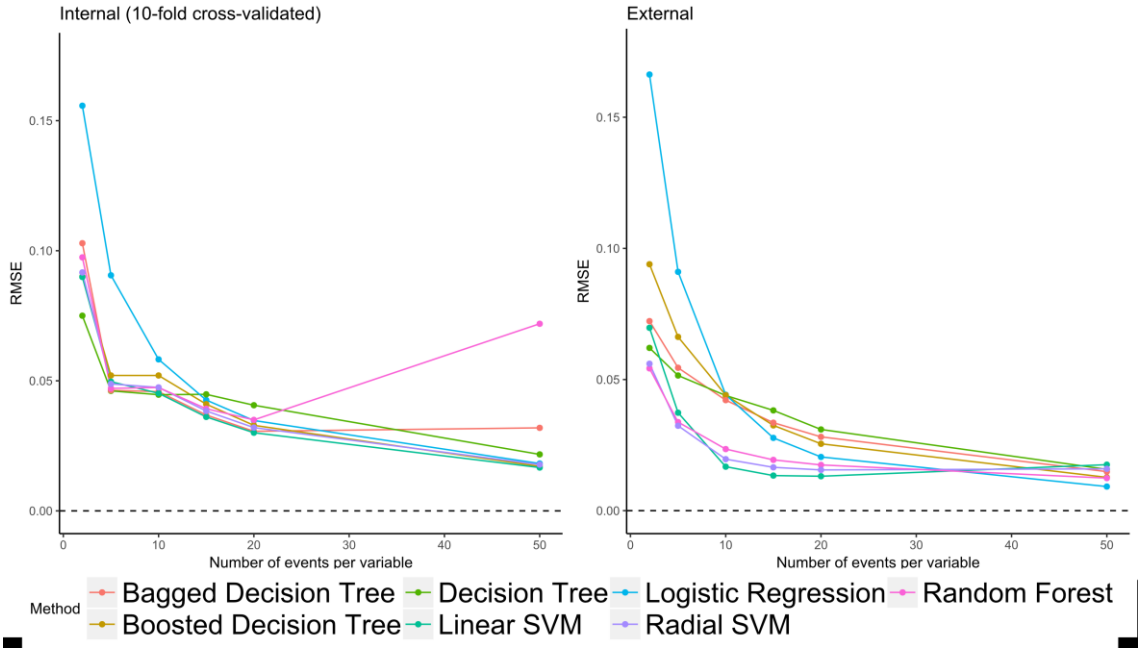


Figure 4.23 RMSE of AUROCs produced using various sample sizes compared to using full internal dataset to development models for each method

Figure 4.24 shows that the RMSE of Brier scores typically decreased with increased numbers of EPV. Although this was not the case for the linear and radial SVMs, with the internal RMSEs being around 0.035 when the number of EPV was 10 or more. The random forest produced a similar pattern for the external RMSEs, with values being around 0.02 when the number of EPV was 10 or more. The logistic regression produced the smallest RMSEs both internally and externally once the number of EPV was 15 or more.

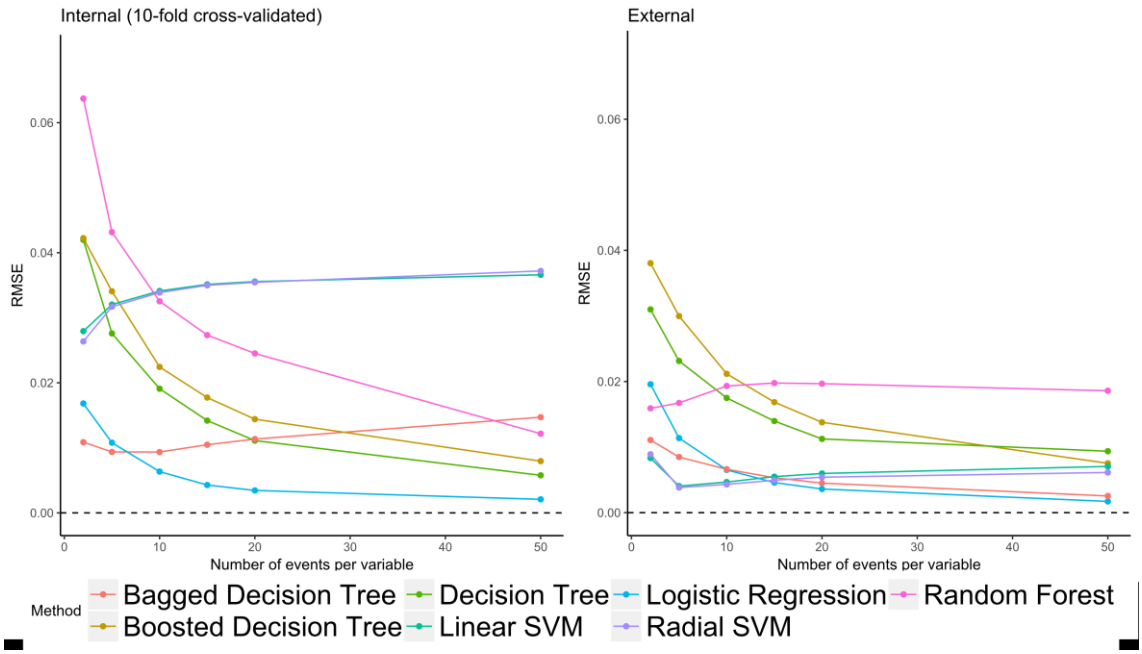


Figure 4.24 RMSE of Brier scores produced using various sample sizes compared to using full internal dataset to development models for each method

Figure 4.25 demonstrates that the majority of simulations with two EPV had external AUROCs not within 10% of their corresponding internal AUROC. The sizable proportion of logistic regression models that have an external AUROC within 1% of the internal AUROC are mainly due to no variables being added to the logistic regression model in these cases. The logistic regression, boosted decision tree, bagged decision tree, linear SVM and radial SVM displayed the expected pattern of increased similarity of the external AUROC to the internal AUROC. In the methods which produced the best AUROCs in the empirical comparison, namely logistic regression, boosted decision tree and linear SVM, the improvement in comparability of internal and external AUROCs was still clear as the number of EPV was increased from 20 to 50.

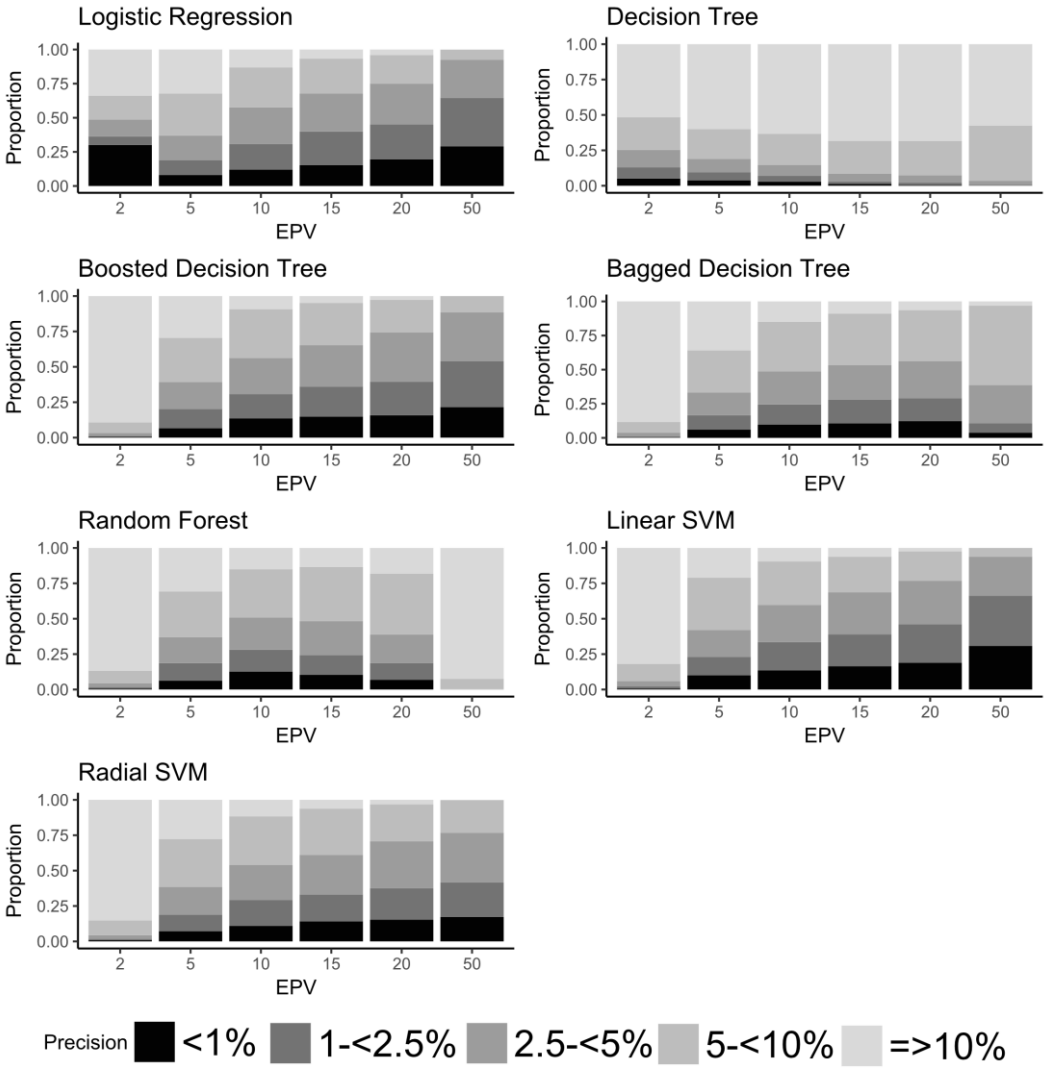


Figure 4.25 Proportion of external AUROCs within 1%, 2.5%, 5%, and 10% of their corresponding cross-validated internal AUROC across methods and sample sizes

4.10 Discussion

4.10.1 Empirical comparison of methods

4.10.1.1 Statistical performance

As stated in Chapter 2 the AUROC was chosen as the metric to assess discrimination since it is a widely used measure in this field, (55,57,59,99). This is the statistic that is of greatest interest to most individuals developing or selecting a RAT to use; as such this is of importance in this comparison. The Brier score was chosen to measure calibration. Although this is not ideal, since it measures both calibration and discrimination together (100,101), when considered alongside the AUROC it gives a good indication of whether models with similar AUROC values have similar levels of calibration. Having scores provides a more reliable comparison of the amount of miscalibration than tests of perfect calibration, such as the Hosmer-Lemeshow, do (73). The Brier score is however affected by the prevalence of the outcome and thus, due to the marginally higher prevalence of the outcome in the external dataset, the Brier scores could be expected to be slightly higher in the external dataset for the complete-case analysis even when AUROCs are equal (180).

The tables displaying the results of each method in the main and sensitivity analyses reported above were selected to give an accurate depiction of how each method performed. Results for changing parameters that were set for all models in a table led, most commonly, to similar results, or occasionally to worse results when deviating from recommendations in the literature or default values in programmes. For this reason and simplicity, the values will be primarily discussed here.

Logistic regression resulted in AUROCs greater than 0.700 for both datasets; for both the RAT which used continuous values of the continuous variables and the RAT which used categorical values of continuous variables. These RATs resulted in AUROC of 0.714 and 0.702 respectively in the external dataset. These values indicate acceptable levels of discrimination for the outcome in this empirical comparison.

The basic decision tree RATs could not match the discrimination of the logistic regression RATs, with discrimination not reaching acceptable levels. Similar poor performance was observed in the sensitivity analysis. The boosted method improved levels of discrimination as the number of trees was increased; with the AUROCs in both datasets reaching 0.70 in the models with 5,000 and 10,000 trees. The Brier scores for these models were very high, around 0.25 in both datasets, which was considerably higher than the outcome index variance. This showed that their calibration is poor, likely due to the weighting of the model in order to gain good levels of sensitivity, which is important in the context of the outcome. Looking at the Brier scores for the basic decision tree models it could be seen that the Brier score increased greatly for each increase in weighting.

The bagged decision tree method produced reasonable levels of discrimination and calibration once the maximum depth of trees allowed was at least five, although with AUROCs being around 0.68 it fell short of matching the metrics achieved using logistic regression. Finally, the random forest method had unsatisfactory discrimination with the highest AUROC in the internal dataset being 0.655 in the main analysis. Further to this the Brier scores were very high indicating bad calibration.

The SVMs with linear kernels gave AUROCs of greater than 0.70 in both dataset for the main analysis, similar to the logistic regression method's performance. On the other hand, the radial SVMs resulted in poor discrimination in the external dataset with AUROCs of 0.628, showing that the AUROCs in the internal dataset were not stable.

The statistical comparison suggests RATs developed using logistic regression method, either using continuous values or categorical values of continuous variables, as well as RATs developed using linear SVM method, give acceptable performance in terms of discrimination and calibration in both internal and external datasets.

4.10.1.2 Comparison of practical issues

This section compares the practical issues of building and using RATs with each of these methods, it focuses particularly on the methods which performed favourably in the statistical comparison. Firstly, the basic decision tree and the

logistic regression RAT with continuous variables grouped could be implemented on a paper platform using only pen and paper (51,110). In contrast the logistic regression RAT with continuous variables kept continuous, with interactions and quadratic terms included, would require an electronic platform for individuals to complete it since the calculations are more difficult. RATs built using boosted decision trees, bagged decision trees, random forests, linear SVMs or radial SVMs would all also require an electronic platform such as a phone app or website in order to allow individuals to calculate their risk.

The paper-based risk score derived from logistic regression carries an educational message; as individuals complete it they learn about some of the risk factors for diabetes, including which ones personally place them at risk (51). A RAT created using the basic decision tree also has this potential to inform individuals about their risk factors, showing which combinations in particular affect their risk. RATs developed using electronic-based logistic regression can incorporate messages to tell individuals how much each risk factor adds to the score; yet since the individual is not required to engage with the RAT they might not take in this message as much as if they had completed a paper-based RAT. The other methods produce RATs which are more like ‘black holes’, where the information about individuals is inputted into an electronic system and only the risk score is reported back, this means individuals cannot see how their score has been calculated and the only way information on the RATs can be given is by generalising the model.

Logistic regression and the basic decision tree are the least computationally intensive of the methods with the other methods, such as radial SVM or random forest, often having much longer run times on normal computer systems. This may make them less attractive to some developers to fit. It should also be considered that these more complex calculations will need to be incorporated into whichever electronic platform they are implemented on.

The comparison of issues other than statistical performance advocates either the logistic regression method or basic decision tree method for developing RATs if statistical performance of all methods was equal.

4.10.1.3 Strengths and weaknesses of this empirical comparison

An empirical comparison was favoured over a simulation study due to the complexity of the causes of abnormal glucose, meaning the assumptions required for a simulation study would be difficult to sensibly choose. The empirical comparison included the logistic regression, decision tree and SVM methods, which are the three methods that have been used in practice to develop RATs for diabetes-related binary outcomes (55,57,59,99). One limitation was several methods were not included in this comparison, such as neural networks, linear discrimination analysis and fuzzy C. In light of the previous empirical studies, in section 4.2, none of these method have been highlight as outperforming all the methods included in this comparison, with them at least being matched by one of the methods included. Additionally, limiting the number of methods meant that the extensions of those included could be studied. For example, the extensions of the decision tree methods which have been established to deal with the issues of stability were included, with recommendations from previous research in this area followed (155-157,160). The two SVM kernels which have performed the best in diabetes and other medical datasets were included (106,126-128). As recommended, the logistic regression RATs developed took into account previous evidence and the health message they would convey (42,57).

Unlike previous empirical comparison studies for binary medical outcomes (124-128), this comparison included an external validation, which is the gold standard of validating performance (40). The high levels of missing data in ADDITION-Leicester meant multiple imputation was carried out which carries its own assumptions which may be incorrect or result in bad performance for a certain method. However, since the external dataset had good levels of complete data this meant that multiple imputation was not required in the external dataset and thus the assumptions made in the multiple imputation were also checked. Furthermore, the sensitivity analysis included also provided a check that none of the methods were underperforming due to logistic assumptions made in the multiple imputation. Using a temporal validation gave methods the greatest chance to be stable in the external dataset, as RATs for abnormal glucose are often found not to be validated in different geographical settings. However, the comparison would have been strengthened by using several internal and

corresponding external datasets from a variety of geographical settings. Finally, the performance of methods depends on the nature of relationship between the outcome and the variables used to model, therefore the results of this empirical comparison are likely to be limited to binary glucose outcomes; hence caution should be taken if using these results to select a method for use in a different medical outcome.

4.10.2 Resampling study on the effect of sample size on the performance of methods

Using small numbers of EPV is likely to lead to RATs giving unreliable estimates of predictive performance. The resampling study in section 4.9 assessed the extent to which this occurred for each of the methods considered in the empirical comparison in this chapter. However, it should be noted that the required number of EPV differs between datasets, with factors such as the prevalence of binary risk factors affecting the reliability of models (181), and thus the results found are somewhat limited to the situation considered. Nevertheless, the study adds to the knowledge of previous resampling studies for cross-sectional outcomes by including several methods and focusing on predictive performance.

Selecting the models which were among the best performing from the empirical comparison for each method allowed the effect of the number of EPV to be assessed in the model, which this chapter recommends for each method. For example, the boosted decision trees built in the resampling study had 5,000 trees; as using 5,000 in the empirical comparison clearly outperformed using 1,000 but produced very similar performance to the model built with 10,000. Since a fully automatic technique was required for logistic regression, the human input part of the method used in the empirical comparison was removed. The number of EPV when developing the models using the full internal dataset was 153.4. This is significantly higher than the currently proposed required sample sizes allowing models with a broad range of numbers of EPV, two to 50, to be compared to their corresponding models developed using 153.4 EPV. The STAR dataset used for external validation of these models had 91 EPV.

The results show the AUROCs and Brier scores of the simulations varied greatly when the numbers of EPV were only two or five. The methods which produced reasonable AUROCs in the empirical comparison, logistic regression, boosted decision tree, bagged decision tree and linear SVM, all had noticeable reductions in the variability of both their internal and external AUROCs as the number of EPV increased throughout the range of values assessed. On the other hand, this pattern was not clear for the random forest simulations, which appeared to have similar external AUROCs for 10, 15 and 20 EPV, or the basic

decision tree method, which produced alike internal AUROCs for 10, 15 and 20 EPV. These two methods had poor levels of discrimination in the empirical comparison and thus the lack of reduction in variability may indicate these methods have an issue with stability. This is certainly true of the random forest which produced apparent AUROCs ≥0.9 in the samples it was developed on in the empirical comparison.

For the majority of methods the variability of the Brier scores reduced throughout the range of the number of EPV assessed; with mean bias decreasing also, apart from for the boosted decision tree which had increasing mean bias. The boosted decision tree produced very high Brier scores regardless of the number of EPV and thus had poor calibration. The mean biases for the linear and radial SVMs in the internal dataset indicate the Brier scores yielded in the empirical comparison may be an overestimate. Although this does not appear to have been repeated in the external dataset, with the absolute mean bias being less than 5%.

For the three methods that produced good levels of discrimination in the empirical comparison, logistic regression, linear SVM and boosted decision tree, the reproducibility of the internal AUROC in the external data continues to improve with increased EPV. Considering the logistic regression for example, medians of 4.2%, 2.9% and 1.7% were seen for 10, 20 and 50 EPV; with 57.6%, 74.9% and 92.5% of simulations having a difference of less than 5% for 10, 20 and 50 EPV respectively. This pattern was also evident for the radial SVM and bagged decision tree; however the random forest and decision tree did not clearly show this increase in reproducibility, adding support to the earlier argument that these methods are not reliable.

The results of this resampling study highlight that for the methods producing good AUROC and/or Brier scores in the empirical comparison, increasing the number of EPV reduces the variability of these metrics and increases the reliability. The results of this resampling study indicate that RATs built using the methods advocated in the empirical comparison in this chapter have noticeable improvements in reliability and performance when the number of EPV is increased from 10 to 20. For that reason the results advocate at least 20 EPV

should be used in datasets similar to those studied in this resampling study. However, the improvement seen in these models when the number of EPV is increased from 20 to 50 highlights a need for careful consideration of the number of candidate variables in light of previous evidence if the number of EPV is significantly below 50.

4.11 Conclusion and implications for this thesis

The empirical comparison advocates using logistic regression and linear SVMs as methods which provide acceptable statistical performance for detecting those with NDH or undiagnosed T2DM both in internal and external datasets. As these methods provide similar statistical metrics, the practical issues around implementing them should be taken into account. Since RATs based on the linear SVM method do not give an easily comprehendible educational message in the way in which RATs derived using logistic regression can, the logistic regression method is favourable in practice.

Using logistic regression with continuous covariables may produce a RAT with marginally better statistical performance in terms of AUROC values than seen when using logistic regression with continuous variables categorised. However, a RAT developed using logistic regression with continuous variables categorised conveys an educational message more clearly and can be used in a number of formats. For these reasons both are used to develop a RAT for NDH or undiagnosed T2DM defined by HbA1c rather than OGTT. This work is detailed in Chapter 6, assessing the need to update the self-assessment risk score in light of the increased use of HbA1c to measure blood glucose in practice. The newly developed RATs are tested against the existing diabetes self-assessment risk score, the LSA, as well as against one another in an external dataset before a recommendation is made about which risk score should be used in practice.

The resampling study supports using at least 20 EPV, though the continued improvement in reliability and performance when the number of EPV are increased to 50 emphasises the need to carefully consider the evidence for including each candidate variable when the number of EPV is between 20 and 50.

Chapter 5: Chain event graphs for developing risk assessment tools using cross-sectional data

5.1 Chapter Outline

This chapter utilises the novel method of chain event graphs (CEGs) to develop a risk assessment tool (RAT) for the outcome of non-diabetic hyperglycaemia (NDH) or undiagnosed Type 2 diabetes mellitus (T2DM) in a cross-sectional dataset. The chapter establishes the technique of CEG-based RATs for cross-sectional outcomes by overcoming the issues that arise from utilising the CEG method for this purpose. Additionally, it assesses the performance of this method in an external dataset for the outcome of interest of this thesis allowing a comparison with the established methods assessed in the previous chapter.

The work in this chapter has been:

- Orally presented:
  Barber SR, Smith JQ, Barclay LM, Bodicoat DH, Davies MJ, Khunti K, Gray, LJ. 'Utilising the Chain Event Graph method to produce a risk score: evaluating the discrimination in detecting a binary diabetes outcome.' At: 36th Annual Conference of the International Society for Clinical Biostatistics. Utrecht, Netherlands. 23rd-27th August 2015 (C46.I)

5.2 Introduction

CEGs are a type of model based on event trees which consider the numerous combinations of categories of each variable and how these relate to one another, simplifying these relationships in the model where this does not lead to information loss. Thus they may produce a RAT which discriminates better than one developed using logistic regression or a decision tree, as they might find useful combinations of risk factors which current commonly used methods could miss. In this chapter, the method is implemented to develop a RAT using the same dataset (ADDITION-Leicester), outcome and explanatory variables as the Leicester Self-Assessment (LSA) score used in order to compare this method to the others included in the comparison in the previous chapter.

CEGs are derived from probability trees by combining nodes in the tree which have the same associated conditional probabilities (182). The nodes of the tree denote different states and the edges express events. CEG models are described by conditional independence statements, each node and the edges emanating from it have a corresponding conditional independence statement which gives the probabilities of the different events occurring from that state regardless of the categories taken for other variables in the tree. CEGs are an extension of discrete Bayesian Networks to allow for asymmetric dependence structures between the variables in the graphical model. Section 5.2.1 uses an illustrative example to explain how CEGs can be derived from a probability tree and section 5.2.2 describes how a RAT can be yielded from a CEG.

5.2.1 Illustrative example

N. B. This example is fictional and has been chosen to explain the method.

Consider the probability tree given in Figure 5.1. The variables ethnicity, family history of diabetes, body mass index (BMI) and diabetes status have been added to the tree. Shafer et al. notes adding variables into trees in the order that they naturally occur is logical and helps to capture the relationship between the variables (183). In the case of this example, consistent with the context in this chapter, the data is cross-sectional so the logical order needed careful consideration and is not completely clear. The order in this example was

chosen to reflect the likely order in which variables present themselves in the real-world. Firstly, an individual's ethnicity is determined at birth, then family history of diabetes may be present already or may occur in the years that follow, next an individual's current BMI can be influenced by their lifestyle in the last few years and finally diabetes is added since these individuals are not known to have diabetes but may have developed it undetected recently. Another choice that needs to be made before a CEG can be built are the categories any continuous variable will take. In this example BMI needed to be categorised and has been categorised in low, average and high.

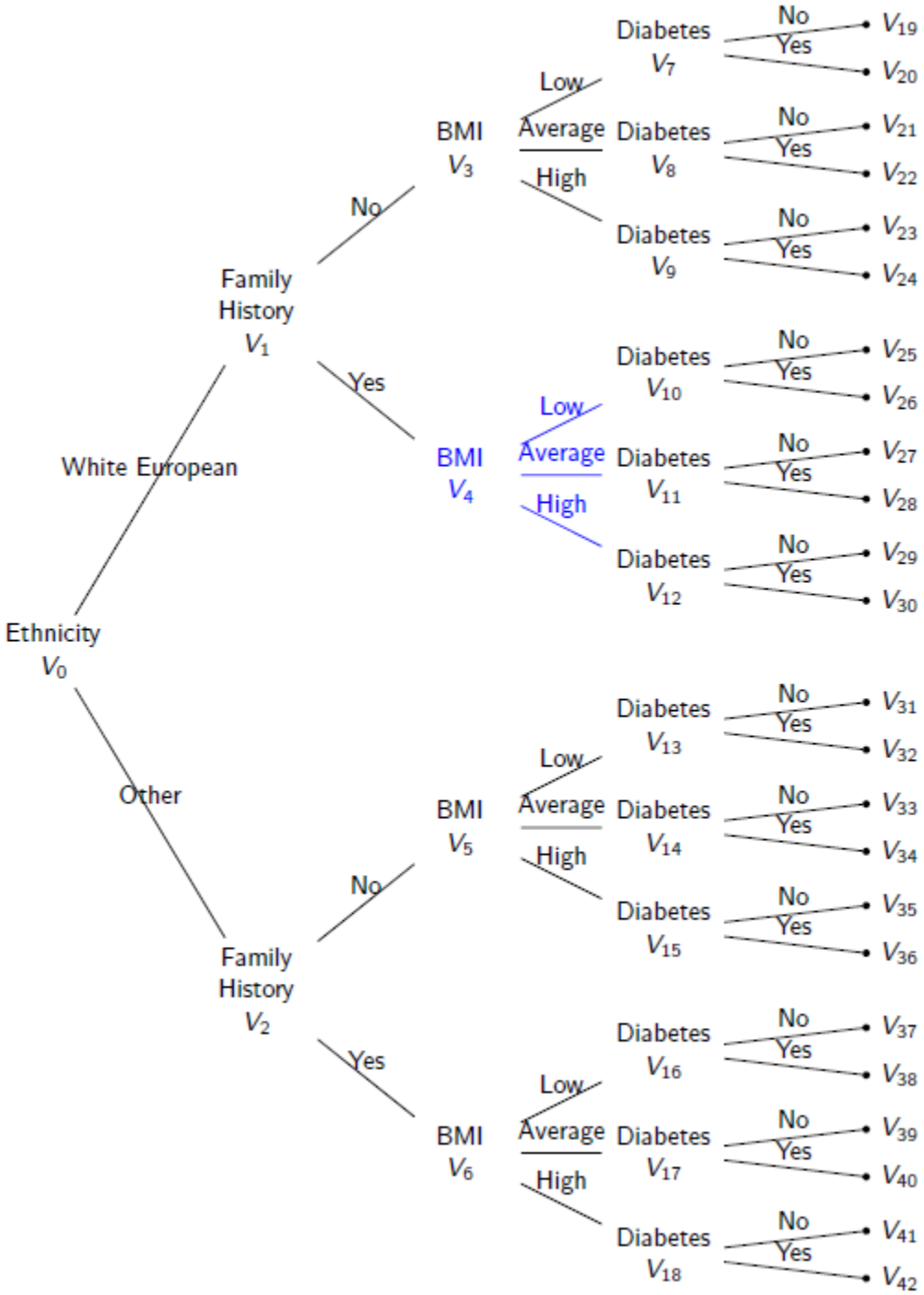


Figure 5.1 Example probability tree for relationship between ethnicity, family history of diabetes, BMI and diabetes status

V_4 and its floret is highlighted

The nodes, known as **situations**, of an event tree, $\mathbf{V}_{i's}$, detail the categories earlier variables in the tree have taken (182). For example an individual being in situation $\mathbf{V_4}$ means that they are a white European with a family history of diabetes. It should be noted, an individual being in situation $\mathbf{V_0}$ just means that they are part of the population of interest. The edges stemming from a situation give the categories that the variable being considered in that situation could take and the probability of each of those categories occurring given the situation, collectively the edges emanating from one situation are known as the **floret** associated with that situation. In the example in Figure 5.1 the floret associated with $\mathbf{V_4}$ gives the probability that an individual will have low, average or high BMI given the fact they are white European with a family history of diabetes.

A CEG model can be yielded from a probability tree by merging situations which have the same associated probabilities in their florets into the same **stage** (182). In order for two situations to be merged into a **stage** the categories in their florets must be exactly the same and the probabilities of each of these categories occurring must be the same for every category. For example, if $\mathbf{V_4}$ and $\mathbf{V_5}$ were merged into a single stage this would mean that the probabilities of having low BMI, average BMI or high BMI given being white Europeans with a family history of diabetes are equal to the probabilities of having low BMI, average BMI or high BMI given being of other ethnicity without a family history of diabetes. Figure 5.2 displays the CEG of the probability tree displayed in Figure 5.1 with $\mathbf{V_4}$ and $\mathbf{V_5}$ merged into a single stage.

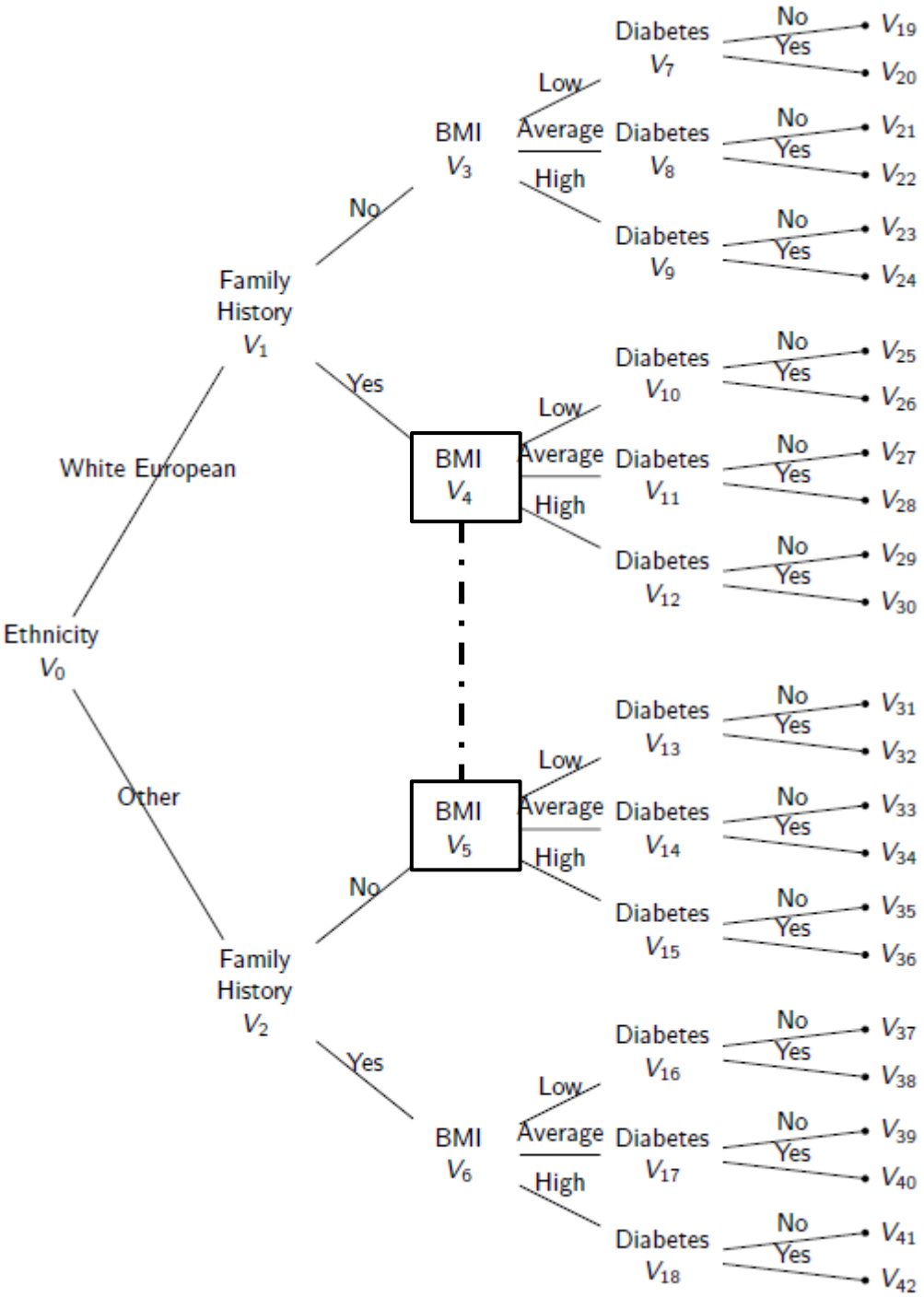


Figure 5.2 Example CEG for relationship between ethnicity, family history of diabetes, BMI and diabetes status with V_4 and V_5 in the same stage

The boxes around V_4 and V_5 and the dotted line linking them indicate that they are in the same stage.

If situations not only have the same associated probabilities in their florets but for their whole subtrees (the trees which originate from each of the situations) then they are said to be in the same **position**, and are displayed as a single node in the CEG (182). More explicitly for two situations to be merged into a **position,** their subtrees must have the same shape with the same categories along each edge. Furthermore each of the matching edges must have the same probabilities as one another. Figure 5.3 displayed the CEG of the probability tree displayed in Figure 5.1 with $\mathbf{V_4}$ and the node for the pathway of other ethnicity followed by no family history merged into a single stage.

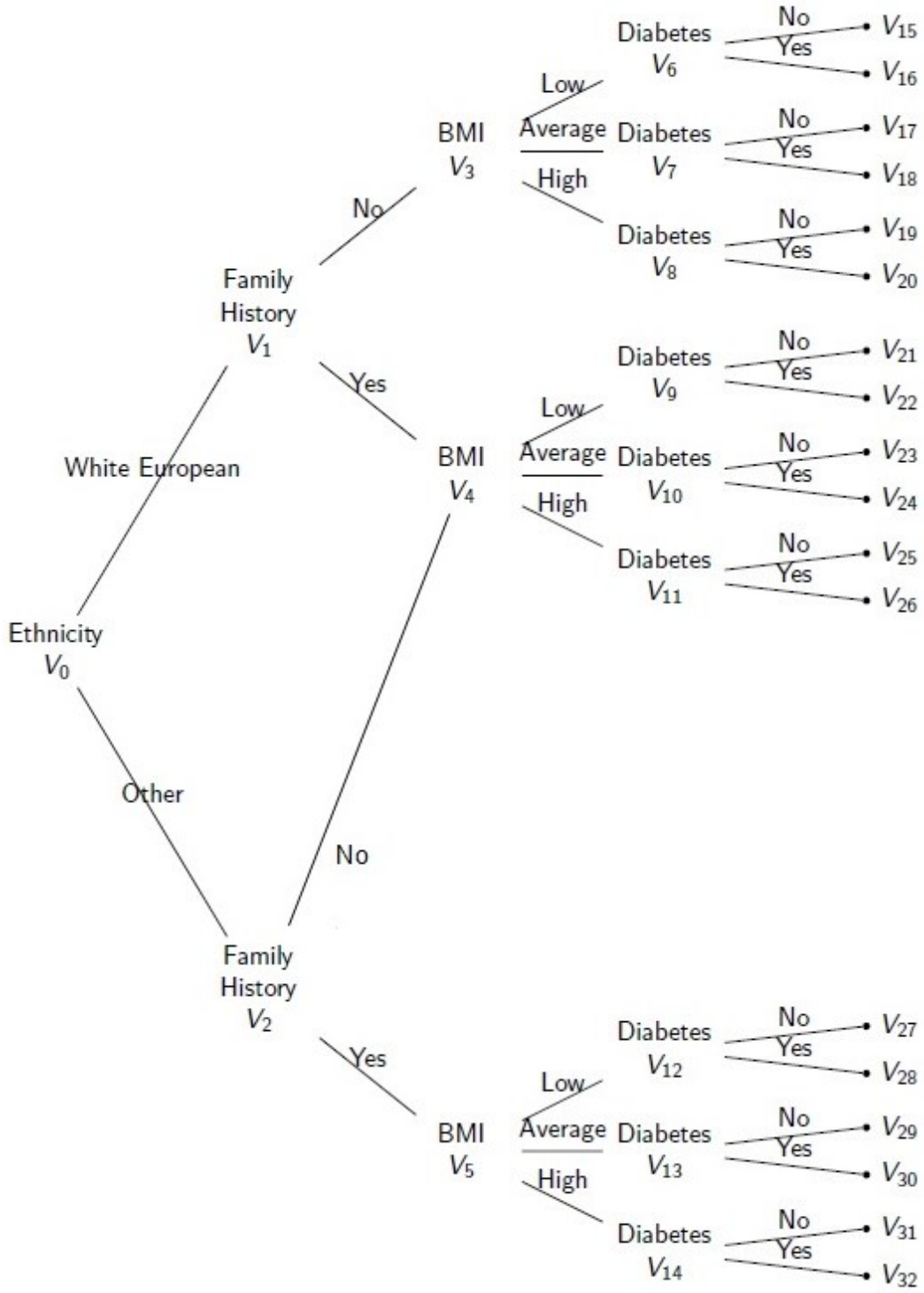


Figure 5.3 Example CEG for relationship between ethnicity, family history of diabetes, BMI and diabetes status

White European individuals with a family history of diabetes are in the same position as non-white European individuals without a family history of diabetes.

In order to decide which situations should be merged a Bayesian approach is taken with numerous CEGs having their posterior likelihood scored and the maximum a posteriori (MAP) model being chosen (184). As CEGs are comprised of several conditional independences, the likelihood of a CEG separates into the product of the likelihoods of each floret being observed (177). Since the data observed at each floret follows a multinomial distribution, giving each conditional probability vector a Dirichlet prior yields the marginal likelihood shown in equation (5.1). Where u specifies the stages of the CEG, n specifies the edges departing from each stage, α_{un} the parameters of the Dirichlet priors describing the probability of having the outcome indexed by n at stage u and x_{un} the data counts giving the number of samples observed having the outcome indexed by n at stage u (185).

$$\prod_u \frac{\Gamma(\sum_n \alpha_{un})}{\Gamma(\sum_n(\alpha_{un}+x_{un}))} \prod_n \frac{\Gamma(\alpha_{un}+x_{un})}{\Gamma(\alpha_{un})} \tag{5.1}$$

As the Dirichlet priors follow on from one another it makes sense to think of the hyperparameters for the priors as dummy counts running through the root-to-sink paths (185). Using non-informative Dirichlet priors would evenly divide a number between the hyperparameters at each split of the tree starting with the root, V_0. For each prior only the number allocated to that path would be available to be evenly split between the hyperparameters.

Under the assumption that the CEG models are a priori equally likely, the different CEGs can be scored and compared using their marginally likelihood alone (186). The Bayesian agglomerative hierarchical clustering (AHC) algorithm does not search all the possible CEGs, as an exhaustive search would; but instead it is a greedy search, in that it chooses the best CEG on each iteration of the algorithm and thus hopes to choose the best CEG out of all possible CEGs (184). Using the AHC algorithm allows CEG models to be selected from event trees with large numbers of situations in reasonable computational time; therefore this is the algorithm that was used when selecting the optimal CEG model from an event tree in this chapter.

The steps of the AHC algorithm are as follows:

1. Start with the CEG, $\mathbf{C_0}$, with the finest partition into stages (where every situation in the original tree is an individual stage)
2. Then combine each of the pairs of nodes which can be combined in turn (scoring each of these CEGs)
3. The CEG (from step 2) with the highest score is selected as $\mathbf{C_1}$ and the pair of nodes which were merged become a single stage
4. This process is repeated until the CEG with the coarsest partition (where every situation in the original tree that can merged has been), $\mathbf{C_n}$, is reached
5. Select CEG of $\mathbf{C_0}$, $\mathbf{C_1}$,.... , $\mathbf{C_n}$ with the highest score

Hypothetical Table 5.1 is an illustration of how the tree in Figure 5.1 may be merged using the AHC algorithm.

Table 5.1 Illustrative example of possible merging of situations at each iteration of AHC algorithm

CEG	Stages merged at iteration
C_0	N/A
C_1	{V_4, V_5}
C_2	{V_{12}, V_{17}}
C_3	{V_{10}, V_{13}}
C_4	{V_4, V_5, V_6}
C_5	{V_{12}, V_{17}, V_{18}}
C_6	{V_7, V_8}
C_7	{V_9, V_{10}, V_{13}}
C_8	{V_{14}, V_{15}}
C_9	{ V_{11}, V_{16} }
C_{10}	{V_{12}, V_{14}, V_{15}, V_{17}, V_{18}}
C_{11}	{V_9, V_{10}, V_{11}, V_{13}, V_{16}}
C_{12}	{V_9, V_{10}, V_{11}, V_{12}, V_{13}, V_{14}, V_{15}, V_{16}, V_{17}, V_{18}}
C_{13}	{V_1, V_2}
C_{14}	{V_7, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}, V_{14}, V_{15}, V_{16}, V_{17}, V_{18}}
C_{15}	{V_3, V_4, V_5, V_6}

Above it can be seen that $\mathbf{V_4}$ and $\mathbf{V_5}$ were merged during the first iteration of the algorithm. Suppose in this example, $\mathbf{C_{11}}$ was the MAP model, the model with the highest posterior likelihood, then the algorithm would chose this as the optimal model. $\mathbf{C_{11}}$ is displayed in Figure 5.4, the positions in the CEG are denoted, $\mathbf{W_i}$.

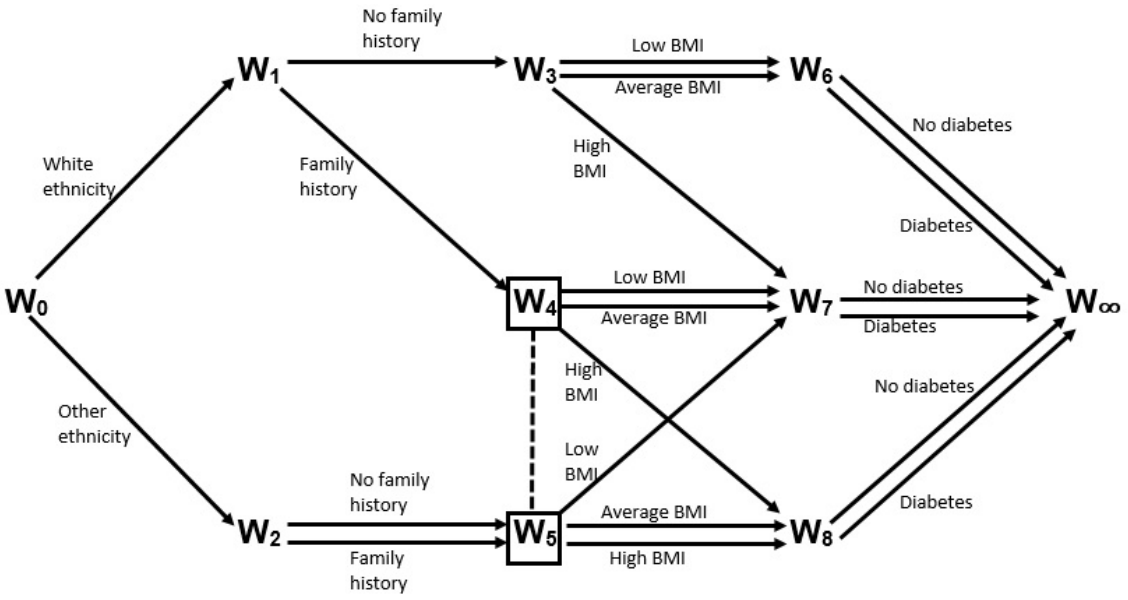


Figure 5.4 C_{11}, the MAP CEG chosen by the AHC in illustrative example.

The boxes around W_4 and W_5 and the line linking them shows that they are in the same stage but different positions.

In C_{11}, displayed in Figure 5.4, the situations $\mathbf{V_5}$ and $\mathbf{V_6}$ have been merged into the same position, $\mathbf{W_5}$. This means individuals not of white European ethnicity with a family history of diabetes have the same probabilities of the different BMI categories and having diabetes or not (given their BMI category) as individuals not of white European ethnicity without a family history of diabetes.
$\mathbf{V_4}$ has been merged into the same stage as $\mathbf{V_5}$ and $\mathbf{V_6}$, as shown in Figure 5.4 by the dotted line between the positions $\mathbf{W_4}$ and $\mathbf{W_5}$. This means the probabilities of the different BMI categories are the same for white European with a family history as for individuals of other ethnicity. However they are not in the same position, because white Europeans with family history of diabetes and average BMI have a different probability of having diabetes to individuals of other ethnicity with average BMI (as shown on Figure 5.4 by the former going to $\mathbf{W_7}$ and the latter to $\mathbf{W_8}$).

5.2.2 How can CEGs be utilised to develop risk assessment tools?

A CEG for the relationship between the risk factors and outcome of interest can be developed. By putting the outcome of interest as the final variable in the event tree (probability tree) from which the CEG is derived, the chosen CEG will contain a number of stages which give the probability of having or developing the outcome of interest given the risk factors of an individual. As a result the chosen CEG model can be used as a RAT, in a similar way to a decision tree (chosen by a classification and regression tree (CART) algorithm for example), by taking the probability of the outcome of interest for an individual as their risk score. However, CEGs account for all interactions of exploratory variables meaning they could find more effective combinations which methods like CART may miss. Continuous risk factors have to be categorised to allow them to be entered into the model.

How a CEG-based RAT could be used in practice depends on the CEG model chosen; though it is likely that the method may produce RATs which are not suitable to be paper based, instead needing to be web/app based because of the different options for different groups. On the other hand a CEG-based RAT may mean that not every individual will need to answer a question on every risk factor a risk score is based on, thereby reducing participants' burden.

5.3 Methods and results of implementing CEG to develop a risk assessment tool

The ADDITION-Leicester dataset was used to develop a RAT based on a CEG. As this was a novel work, the method was altered a couple of times based on shortcomings of the methodology found in this internal dataset. Once these issues had been overcome, the external validity of the technique was assessed using the Screening Those at Risk (STAR) dataset. The outcome was NDH (impaired glucose regulation (IGR)) or undiagnosed T2DM, the same outcome used to develop the LSA score. Additionally, the final RAT developed used the same risk factors as the LSA enabling this method to be compared with the other methods included in the comparison in the previous chapter, since the same internal and external datasets were also used. R studio was used to carry out the work in this chapter, as no package was available for choosing CEGs by the AHC algorithm at the time this work was carried out.

5.3.1 Initial method and results in white Europeans with no family history of diabetes

As this was a novel application of the technique I envisaged that using it on real data would raise methodology issues about how to implement the method in a real world context. I therefore decided to initially build CEG models for a subset of the data, white Europeans with no family history of diabetes, using only four of the risk factors included in the LSA score before dealing with any methodological issues. I then went on to apply the method to develop a RAT on the whole dataset using all seven risk factors included in the LSA score. The 3,368 individuals aged 40- 75 years old in the ADDITION-Leicester dataset with values for all four risk factors and the outcome variable were used for this analysis. 14.7% of these individuals had the outcome of NDH or undiagnosed T2DM.

The order of the variables in the event tree, from which CEGs were derived, and the categories used to split the tree into situations were as follows:

- X_1 : Age category (40-49yrs, 50-59 yrs, 60-69 yrs, 70+ yrs)
- X_2: Waist category (<90 cm, 90-99 cm, 100-109 cm, 110 cm+)
- X_3: BMI category (<25, 25-29, 30-34, 35+)
- X_4: Hypertension status (yes or no)

- X_5: Abnormal glucose (yes or no)

All possible situations were accounted for in this tree. Each conditional probability vector was given a non-informative (uniform) Dirichlet prior distribution. The prior Dirichlet distributions were given an equivalent sample size of 4.The MAP model found using the AHC search algorithm was selected as the CEG to calculate the risk scores from.

The CEG model chosen had 21 stages, seven of which were probability vectors for the diabetes status of interest. This means the method grouped the 128 possible combinations of categories of the four risk factors into seven groups. The probability of having the diabetes status of interest given the stage (determined by the values of the four risk factors) is shown for each of the seven stages in Table 5.2. For example, as Table 5.3 shows individuals aged over 70 years old with a waist of 100-109 cm, BMI of 30-34 Kg/m^2 and no hypertension are in the W_{17} and therefore, as displayed in Table 5.2, their chance of having the diabetes status of interest is 0.434.

Table 5.2 Conditional probability of NDH or undiagnosed T2DM given stage in initial CEG for white European individuals with no family history of diabetes from ADDITION-Leicester dataset (n=3,368)

Stage	Probability of NDH or undiagnosed T2DM given stage in CEG	Proportion of internal dataset (%)
W_{14}	0.037	16.18
W_{15}	0.002	2.94
W_{16}	0.226	21.94
W_{17}	0.434	6.29
W_{18}	0.091	25.42
W_{19}	0.147	27.11
W_{20}	0.992	0.12

Taking the probability of having the diabetes status of interest, from the conditional probability vectors produced by the CEG, as the risk score for individuals gives a score with good discrimination. The area under the receiver operator curve (AUROC) of this score is 0.712 (95% confidence interval (CI):

0.689-0.735). However, despite this good performance in terms of discrimination, Table 5.3 demonstrates the problem encountered with sparse data for several pathways (combinations of variable categories). There were numerous pathways in which there were less than 10 individuals; these pathways have been given a risk score under this model based on very little data and, in several cases, no data at all. Clearly this was a big methodological flaw that needed to be addressed. Looking at the pathways which had no individuals at all, it is apparent this was to be expected in some instances, for example all pathways which had the combination of the smallest waist grouping, less than 90 cm, and the largest BMI grouping, greater than 35 Kg/m^2, had no observations.

Table 5.3 Number of cases and non-cases for each situation in W_{17} in the initial CEG for white Europeans with no family history in ADDITION-Leicester

Age	Waist	BMI	Hypertension	Cases	Non-cases
40-49	<90	35+	No	0	0
40-49	<90	35+	Yes	0	0
40-49	100-109	<25	No	0	0
40-49	100-109	<25	Yes	0	0
40-49	110+	<25	No	0	0
40-49	110+	<25	Yes	0	0
40-49	110+	25-29	Yes	0	0
40-49	110+	35+	Yes	6	13
50-59	<90	30-34	Yes	0	0
50-59	<90	35+	No	0	0
50-59	<90	35+	Yes	0	0
50-59	100-109	<25	Yes	0	0
50-59	110+	25-29	Yes	0	0
60-69	<90	35+	No	0	0
60-69	100-109	30-34	Yes	20	34
60-69	100-109	35+	No	3	5
60-69	110+	<25	Yes	0	0
60-69	110+	35+	Yes	23	19
70+	<90	35+	Yes	0	0
70+	90-99	35+	Yes	0	0
70+	100-109	30-34	No	5	9
70+	100-109	30-34	Yes	11	11
70+	100-109	35+	No	2	3
70+	110+	<25	Yes	0	0
70+	110+	25-29	No	0	0
70+	110+	25-29	Yes	3	2
70+	110+	30-34	No	4	6
70+	110+	30-34	Yes	8	11
70+	110+	35+	Yes	7	7

5.3.2 Updated method and results in white Europeans with no family history of diabetes

Two changes were made to the technique to deal with the issues of sparse data identified in the initial analysis. Firstly waist-group specific cut-points were used to define BMI groups instead of the cut-points originally used; meaning that the BMI group an individual was categorised into depended on which waist group an individual was in as well as their BMI measurement. These cut-points used

values which roughly split the individuals in the four waist groups into thirds based on their BMI, in order to allow the BMI variable to add additional information to the correlated waist variable, the variable thus became 'for their waist group does this individual have a low, average or high BMI?'. The cut-points used are showed in the Table 5.4.

Table 5.4 Waist-specific BMI cut-points for groupings BMI as low, average or high

Waist (cm)	**BMI cut-points for different waist groups**		
	Low	Average	High
< 90	<23	≥23 & <26	≥26
>89 & <100	<26	≥26 & <29	≥29
>99 & <110	<29	≥29 & <32	≥32
>109	<33	≥33 & <37	≥37

The second change to the method was to implement a stopping rule which meant the original probability tree did not allow any nodes to split into a group of nodes that contained a node with less than 10 individuals in it. The stopping rule implemented will be referred to as Stopping rule 1 in this chapter. Stopping rule 1 was as follows (detailed in italics): *The situations with less than 10 individuals are to be merged conservatively, meaning if for example at one situation splitting for age resulted in the 50-59yr old category having less than 10 individuals this would be merged with the category above (the higher risk category, in this case 60-69yr old) as it would overestimate the risk for those in the 50-59yr old category rather than underestimate. This was not done in the case where the top risk category had sparse data; in this case it had to be merged with the category below. If there was more than one category within a variable with less than 10 individuals the one with the lowest number would be merged first.*

Again, the 3,368 individuals aged 40- 75 years old in the ADDITION-Leicester dataset with values for all four risk factors and the outcome variable were used for this analysis.

The order of the variables in the event tree, from which CEGs were derived, and the categories used to split the tree into situations were as follows:

- X_1 : Age category (40-49 years, 50-59 years, 60-69 years, 70+ years)
- X_2: Waist category (<90 cm, 90-99 cm, 100-109 cm, 110 cm+)
- X_3: BMI category (low, average, high)
- X_4: Hypertension status (yes or no)
- X_5: Abnormal glucose (yes or no)

Each conditional probability vector was given a non-informative (uniform) Dirichlet prior distribution. The prior Dirichlet distributions were given an equivalent sample size of 4. These priors along with the data were merged according to Stopping rule 1 to give the event tree on which the AHC search algorithm was carried out to find the MAP model.

These solutions were reasonably successful with only eight pathways having less than 10 cases and controls showing that the approach with the BMI groups depending on waist group is sensible. All pathways which required the stopping rule occurred when splitting for the last risk factor, hypertension, and thus were merged/not split at this stage. A similar level of discrimination was seen to the initial CEG-based RAT with an AUROC of 0.7085 (95% CI: 0.685-0.732). The CEG model chosen had 15 stages, five of which were probability vectors for the diabetes status of interest. The probability of having the diabetes status of interest given the stage (determined by the values of the four risk factors) is shown for each of the five stages in Table 5.5. The chosen CEG was very complex and therefore it is difficult to depict the model or portray the individuals who were put into each of the five stages giving the risk for the outcome of interest.

Table 5.5 Conditional Probability of NDH or undiagnosed diabetes given stage in updated CEG for white European individuals with no family history of diabetes from ADDITION-Leicester dataset (n=3,368)

Stage	Probability of NDH or undiagnosed diabetes given stage in CEG	Proportion of internal dataset (%)
W_{10}	0.027	13.42
W_{11}	0.079	29.63
W_{12}	0.149	32.93
W_{13}	0.423	7.66
W_{14}	0.236	16.36

5.3.3 Issues with implementing updated method with seven risk factors in all 40- 75 year olds

The updated approach with waist-specific BMI groups and Stopping rule 1 was implemented on the 6,101 individuals aged 40- 75 years old in the ADDITION-Leicester dataset with values of the seven risk factors from the LSA score and of the outcome recorded. 17.6% of these individuals had the outcome of NDH or undiagnosed diabetes.

The order of the variables in the event tree and the categories used to split the tree into situations were as follows:

- X_1: Ethnicity (White European or other)
- X_2: Sex (male or female)
- X_3: Family history of diabetes (yes or no)
- X_4: Age category (40-49 years, 50-59 years, 60-69 years, 70+ years)
- X_5: Waist category (<90 cm, 90-99 cm, 100-109 cm, 110 cm+)
- X_6: Waist-specific BMI category (low, average, high)
- X_7: Hypertension status (yes or no)
- X_8: Abnormal glucose (yes or no)

Each conditional probability vector was given a non-informative (uniform) Dirichlet prior distribution. The prior Dirichlet distributions were given an equivalent sample size of 4. These priors along with the data were merged

according to Stopping rule 1 to give the event tree on which the AHC search algorithm was carried out on.

However, implementing stopping rule 1 on this dataset with the model being extended to include seven risk factors proved to be very complex. As was the case with applying this technique to the subset of white Europeans with no family history using only four risk factors, merging was required from the fourth variable splits. Though the increase in the number of risk factors used meant that the number of merges required in this case was very high. Table 5.6 shows there were 267 different situations being the first point at which a pathway had less than 10 individuals. Stopping rule 1 leading to the 768 possible situations individuals may be split into using the seven risk factors being collapsed into 222 (28.9%) situations. This meant risk factors later in the event tree did not add any information for many individuals.

Table 5.6 Number of collapses required when implementing Stopping rule 1 on event tree with seven risk factors in 40- 75 year olds from ADDITION-Leicester (n=6,101)

Variable used to split	Original number of resulting nodes	Nodes NAs due to <10 on this split	Nodes NAs due to <10 on previous split
Age	32	3	0
Waist	128	20	12
BMI	384	110	96
Hypertension	768	134	412

More problematic was the fact that the merges resulting from Stopping rule 1 lead to situations having florets emerging from them with several different topographies, either in terms of the number of categories stemming from the situation or in terms of the groupings of the categories. This was an issue as it made the AHC algorithm very complex and time-consuming to code and the resulting CEG complex to understand. For example implementing Stopping rule 1 meant that the florets for the waist situations could have one of 11 topographies.

5.3.4 Final method and results in all 40- 75 year olds

The final method was developed using the 6,101 individuals aged 40- 75 years old in the ADDITION-Leicester dataset with values of the seven risk factors from the LSA score and of the outcome recorded.

As implementing Stopping rule 1 on the full dataset lead to the technique becoming very complex and time-consuming, the stopping rule for creating the event tree on which the AHC algorithm was implemented was modified. This stopping rule will be referred to as Stopping rule 2. Stopping rule 2 was as follows (detailed in italics): *If splitting a situation into a group of situations using one of the risk factors results in any of the situations having less than 10 individuals then this variable will not be used to split this situation at all. The situation may still be split for later risk factors in the model if this does not result in situations with less than 10 individuals in.*

Clearly this stopping rule would result in the data for the later variables in the model not being utilised to determine many of the individuals' risk of having the outcome of interest. Therefore, as the risk factors are from a cross-sectional dataset they were added into the model in order of their p-value with the outcome of interest, with the variable with the smallest p-value being added first. The order of the variables in the event tree and the categories used to split the tree into situations were as follows:

- X_1: Waist category (<90 cm, 90-99 cm, 100-109 cm, 110 cm+)
- X_2: Age category (40-49 years, 50-59 years, 60-69 years, 70+ years)
- X_3: Hypertension status (yes or no)
- X_4: Waist-specific BMI category (low, average, high)
- X_5: Ethnicity (White European or other)
- X_6: Family history of diabetes (yes or no)
- X_7: Sex (male or female)
- X_8: Abnormal glucose (yes or no)

Each conditional probability vector was given a non-informative (uniform) Dirichlet prior distribution. The prior Dirichlet distributions were given an equivalent sample size of 4. These priors along with the data were merged according to Stopping rule 2 to give the event tree on which the AHC search algorithm was carried out on.

The CEG model chosen had 52 stages, seven of which were probability vectors for the diabetes status of interest. The probability of having the diabetes status

of interest given the stage is shown for each of the seven stages in Table 5.7, along with the proportion of individuals from the internal dataset that were in each stage. The internal discrimination of the RAT produced was good, with an AUROC of 0.720 (95% CI: 0.705-0.736). The Brier score in the internal data was 0.131, with the outcome index variance being 0.145 showing unsurprisingly the predicted probabilities are more accurate than non-informative prediction of the outcome prevalence. As with the earlier CEGs, the model is too complex to display pictorially and thus it is hard to summarise the characteristics of individuals placed into each of the stages for risk of the outcome.

Table 5.7 Conditional Probability of NDH or undiagnosed diabetes given stage in initial CEG for 40- 75 year olds from ADDITION-Leicester dataset (n= 6,101)

Stage	Probability of NDH or undiagnosed diabetes given stage in CEG	Proportion of internal dataset (%)
W_{45}	**0.073**	19.93
W_{46}	**0.024**	6.83
W_{47}	**<0.001**	2.49
W_{48}	**0.136**	31.13
W_{49}	**0.199**	11.46
W_{50}	**0.283**	20.23
W_{51}	**0.475**	7.93

The RAT produced using this method was assessed externally in the STAR dataset. The dataset contained 3,105 individuals aged 40- 75 years old with complete data for the seven risk factors and outcome variable. 19.6% of individuals had the outcome of NDH or undiagnosed diabetes. The AUROC of the RAT in this external dataset was 0.639 (95% CI: 0.615- 0.663). The Brier score was 0.153, with the outcome index variance being 0.158. This shows the predictions have accuracy and calibration worse than seen in the internal dataset but a little better than the non-informative prediction of the outcome prevalence. Figure 5.5 displays the predicted probability of the outcome for each stage against the observed proportion with the outcome in that stage in the external dataset. It can be seen that the four stages with the lowest probability predictions observed higher proportions than predicted in the

external dataset, while the three stages with the highest probability predictions observed lower proportions than predicted.

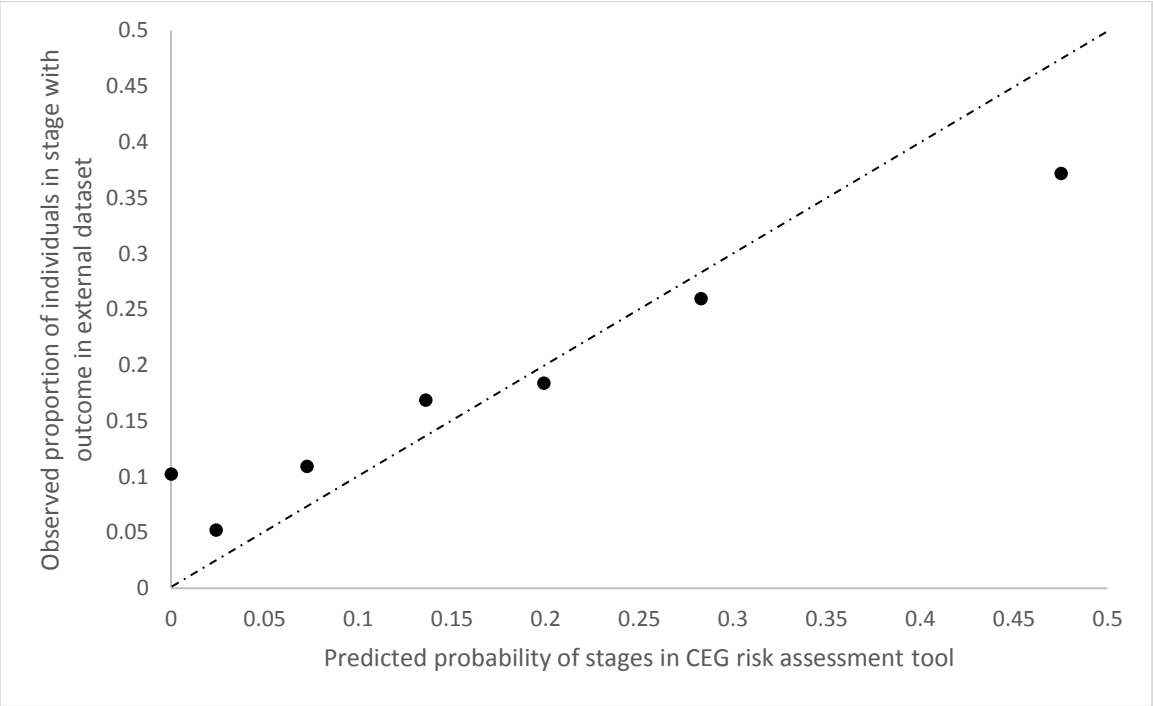


Figure 5.5 Plot of predicted probabilities of outcome (NDH or undiagnosed diabetes) by CEG risk assessment tool's stages against observed proportions of outcome in CEG risk assessment tool's stages in STAR dataset (n=3,105)

5.4 Discussion

This chapter implemented a novel application of the CEG model, using the technique to develop a RAT. The method established performed well in the internal dataset on which it was developed with good levels of discrimination for the outcome of interest, similar to those seen for the LSA score (51). However, the discrimination dropped noticeably in the external dataset and this would need to be addressed before the technique could be recommended for developing RATs to be employed in practice.

The RATs developed using logistic regression and linear SVM in the previous chapter discriminated the outcome comfortably better in the external dataset than the RAT produced using the method established in this chapter. In addition to being statistically outperformed, the CEG model produced was very complex unlike the one produced in the illustrative example in Figure 5.4, with it proving impossible to depict the graph in an eligible form. This means the resulting RAT produced would not give an easy to understand educational message. Such RATs, would also require individuals to calculate their risk using an electronic device.

This analysis was carried out using ADDITION-Leicester as the internal dataset and STAR as the external dataset, this was a strength as it allows the performance of the final RAT established to be compared to the performance of the methods included in the comparison in the previous chapter. The seven risk factors included in the LSA score were used in the final RAT developed using the CEG method in this chapter. The reason for this was that using all the candidate variables that were considered when comparing methods in the previous chapter was not viable due to the high number of them, and thus some preselection of the candidate variables was required. This may have resulted in a variable that may have improved the performance being omitted. On the other hand, using all seven risk factors included in the LSA score may been over complicated the CEG RAT developed. As it may have been possible to achieve a similar performance using fewer risk factors and thus from a simpler RAT developed using the CEG method. A complete-case analysis was used as using multiply imputed data was not computationally viable. This may have led to bias

in the resulting model selected; though, reassuringly the seven risk factors used had low levels of missing data.

The CEG method requires continuous variables, such as age, to be categorised. Categorising variables leads to loss of information given by a variable, with information loss being increased as the number of categories used is reduced (187). The categories selected for continuous risk factors were originally chosen to be the same as those used for the LSA score. However to overcome the issue of sparse data due to the high correlation between BMI and waist circumference, the cut-points for the BMI groups were changed to be waist-group specific. This meant the BMI group was more informative about the difference between individuals given their waist group was already known.

A stopping rule was introduced to ensure situations which had no or little data were not included in the event tree from which the CEG model was developed. The stopping rule had to be simplified to allow the coding of the event tree from which the CEG model was developed to be viable in a reasonable time-frame when the model was extended to use all seven risk factors included in the LSA score. The use of a larger dataset or fewer risk factors could overcome the need for a stopping rule when applying this technique.

The technique established put the risk factors into the model in ascending order of their p-value with the outcome. This was due to the need for a stopping rule meaning later risk factors were less likely to effect the stage which gave the probability of the outcome (the risk score) an individual was allocated by the selected CEG model. Although it is advocated that variables are entered into CEG models in the order which reflects how they would be observed in practice (183). As the RAT developed here uses cross-sectional risk factors it is unclear which order several of the risk factors occur in practice. For this reason the analysis would benefit from various techniques for determining the order of the risk factors being compared. This was not possible at the time the analysis was carried out due to the time-consuming nature of coding the technique, however some software is being developed for CEGs currently which would reduce the burden of building the models required.

Finally, the analysis was limited to the use of non-informative priors. This was in order to allow the method to be established and compared to the other methods, without burdening experts with the time-consuming nature of this task. If the issues with the external validity were resolved, the Bayesian structure of this method would allow expert opinion to be incorporated into RATs developed (188).

5.5 Conclusion and implications

The method of CEGs can be used to develop RATs for cross-sectional outcomes. The method established produced a RAT for the outcome of NDH or undiagnosed diabetes with good internal discrimination, yet this was not observed in the external dataset and therefore does not offer any improvement on the methods currently used to develop RATs. Current advances in the field of CEGs, such as software being developed to reduce the need for coding, will facilitate further exploration of possible adaptions to improve the technique's external performance. Future work could examine the effect of the number of risk factors in the model, the cut-points used to group continuous variables and the order the variables are entered into the model.

Chapter 6: Assessing the impact of HbA1c diagnostic criteria on Leicester Self-Assessment risk score

6.1 Chapter Outline

Since the development of the widely used Leicester Self-Assessment (LSA) risk score, glycated haemoglobin A1c (HbA1c) has been added to the diagnostic criteria for Type 2 diabetes mellitus (T2DM) and many also now use HbA1c to define non-diabetic hyperglycaemia (NDH). Therefore this chapter considers whether the LSA needs to be updated in light of the increased use of HbA1c as the blood test for assessing diabetes status. Two self-assessment risk assessment tools (RATs) for the outcome of abnormal HbA1c are developed; one using the exact values of the continuous variables and giving the probability of the outcome, the other grouping continuous variables and giving a risk score as a whole number. The performance of these RATs in differentiating between individuals with normal and abnormal levels of HbA1c are then assessed and compared to the performance of the existing LSA risk score in an external dataset.

6.2 Introduction

The LSA risk score identifies individuals with a high risk of currently having NDH or undiagnosed T2DM. It is widely used in practice, having been completed over one million times on the Diabetes UK website as well as being in all Boots and Tesco stores across the country. The LSA risk score was developed and externally validated for the outcome of impaired glucose regulation (IGR) or undiagnosed T2DM (51). IGR was defined using either of the two blood glucose measurements taken during an oral glucose tolerance test (OGTT), as detailed in Chapter 1. At the time the LSA was developed and validated an OGTT was the only advocated blood test for diagnosing individuals as having NDH or T2DM (13). In the years since the LSA was developed, HbA1c has been recommended as another blood test which can be used to diagnose diabetes by the World Health Organisation (WHO) (12). Additionally, HbA1c has been widely supported as a measure to determine NDH, although organisations recommend different ranges to define NDH (14,28,29). HbA1c has the advantage of not requiring individuals to fast prior to the test and thus can be scheduled at any time over the day, unlike OGTT (39). This has led to many general practices opting to use HbA1c as their preferred test for classifying glucose intolerance. Although, changing from OGTT to HbA1c testing will result in different individuals being identified as having NDH and undiagnosed T2DM. Studies have found that the two tests classify different but overlapping groups as having undiagnosed T2DM, as displayed in Figure 6.1, with the same being true for NDH (39).

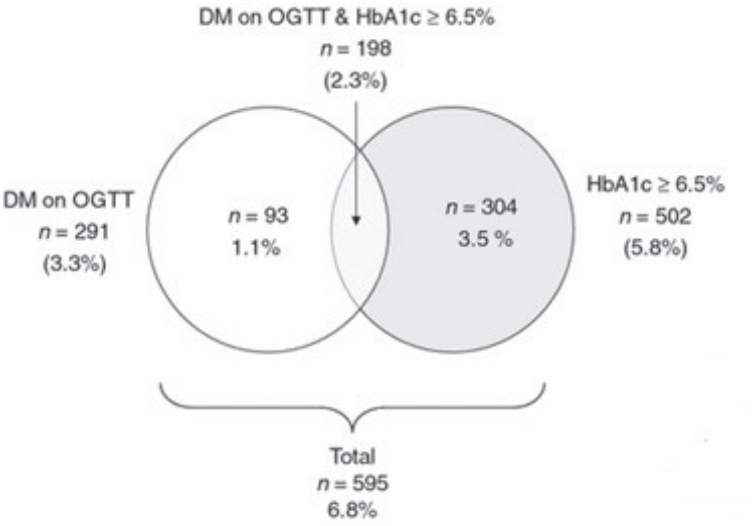


Figure 6.1 Venn diagram of individuals from the Leicester Ethnic Atherosclerosis and Diabetes Risk (LEADER) cohort classified as having undiagnosed T2DM using OGTT and HbA1c, adapted from Mostafa et al. (32)

In light of the increased use of HbA1c in practice, this chapter considers whether a RAT developed for the outcome of abnormal HbA1c performs better in identifying individuals with that outcome in an external dataset than the existing LSA risk score. This chapter reports the development and performance of two self-assessment RATs for the outcome of abnormal HbA1c. The first RAT developed uses exact values of the continuous variables and gives the probability of having an abnormal HbA1c measurement. Keeping the exact values of continuous variables is considered best practice, as categorising such variables results in a loss of information and power (187,189). However, the majority of use of the LSA in practice is using its written version in roadshows or opportunistically. Due to the more complex calculations, this RAT with exact values of continuous variables would have to be implemented using an electronic device, such as an app or website. The need for an electronic device, which may limit its use in clinical practice and for opportunistic screening compared to the existing LSA, will be taken into account when comparing this RAT with the LSA. This RAT is referred to as the electronic RAT throughout this chapter.

The second RAT developed, groups the continuous variables in the same way as the LSA did, leading to a risk score which is a whole number that can be calculated easily without the need for an electronic device. This RAT would be

able to be completed using pen and paper by lay people, meaning it can be compared directly to the LSA risk score since it could be implemented in the same way. This RAT is referred to as the pen and paper RAT throughout this chapter.

The two RATs were developed using the same dataset, ADDITION-Leicester, which was used to develop the LSA risk score. Both the RATs were then compared to the LSA and one another in an external dataset to assess if either should replace the LSA risk score in practice. The two RATs were also compared to one another to determine which RAT should be chosen if it was deemed they both could replace the LSA risk score.

6.3 Methods

The data of 40- 75 year olds in the ADDITION-Leicester dataset was used to develop two RATs, a pen and paper RAT and an electronic RAT, for the outcome of HbA1c ≥6.0%. Logistic regression was used to build both RATs. The following variables were considered for both the RATs: age, sex, ethnicity, body mass index (BMI), waist circumference, first degree family history of T2DM, history of high blood pressure or antihypertensive use, steroid use, smoking status and previous diagnosis of each of the following conditions: high cholesterol, angina, stroke, gestational diabetes and polycystic ovary syndrome. Only 68.0% of individuals aged 40-75 years old in dataset had complete data on all candidate variables as well as the outcome variable, therefore multiple imputation of missing candidate variables was performed using fully conditional specification (FCS) in Stata 13 (146) using the *ice* program.

6.3.1 Multiple imputation

Multiple imputation using *ice* has been explained in detail in Chapter 4. The multiple imputation carried out on the internal data for this chapter used 50 imputations. The FCS specified each variable with missing values a conditional distribution. These used the candidate variables and the following variables: HbA1c, systolic blood pressure, diastolic blood pressure, statins use, alcohol consumption, fruit and vegetable consumption (daily or not), exercise (30 minutes daily or not), surgery for coronary arteriosclerosis, cardiac arrhythmia and history of high blood glucose. Continuous variables which did not follow a normal distribution were log transformed before imputation was carried out and then back transformed along with the imputed values afterwards. Continuous variables which still did not follow a normal distribution after log-transformation were imputed using bootstrapping to overcome this issue.

6.3.2 Development of electronic risk assessment tool

Firstly, backward elimination was carried out starting with the full logistic regression model containing all candidate variables and removing variables until all remaining variables had $p<0.05$. After the automatically selected model had been produced, it was then considered whether any eliminated variables should be added to or replace any of the variables automatically selected due to either previous evidence or the health message resulting from their inclusion. The

internal area under the receiver operator curve (AUROC) of the models considered was calculated to help inform the model selection. Once the linear terms to be included in the electronic RAT were selected, interactions and quadratic terms were considered. Forward selection was used to add interactions and quadratic terms to the linear terms. Due to the large number of interactions tested, the interactions and quadratic terms were added to the linear terms if their p-value was lower than 0.01 and 0.05 respectively. The electronic RAT was simply the logistic regression model of the final model once interactions and quadratic terms had been added.

6.3.3 Development of pen and paper risk assessment tool

Again backward elimination was carried out starting with the full logistic regression model containing all candidate variables and removing variables until all remaining variables had $p<0.05$, the continuous variables were kept continuous for this procedure. Once the variables had been selected using this automated approach, the continuous variables were then grouped and a logistic regression model containing the automatically selected variables but with continuous variables grouped was built. The continuous variables were grouped as follows: age: 40-49 years old, 50-59 years old, 60-69 years old and 70-75 years old; BMI: <25 kg/m^2, 25 kg/m^2 ≤ to <30 kg/m^2, 30 kg/m^2 ≤ to <35 kg/m^2 and ≥35 kg/m^2; waist circumference: less than 90 cm, 90-99 cm, 100-109 cm and 110 cm or more. After the automatically selected variables had been fitted with continuous variables grouped it was then considered whether any of the eliminated variables should be added to or replace any of the variables automatically selected due to either previous evidence or the health message resulting from their inclusion. The internal AUROC of the models considered was calculated to help inform the model selection. Finally the chosen logistic regression model was used to produce a simplified RAT that can be completed using only pen and paper. The RAT allocates a score to an individual by giving them a score for each category they fall into, which is the coefficient of the logistic regression model for that category multiplied by 10 and rounded to the nearest whole number, the sum of the scores for each category is the individual's total risk score.

6.3.4 Comparison of two HbA1C risk assessment tools and LSA risk score

Data of 40- 75 year olds from the Screening Those at Risk (STAR) dataset, which has been detailed in Chapter 2, were used for external validation of the two RATs developed in this chapter and comparison with the LSA risk score. Since the levels of missing data in the STAR dataset were low, 1.9% of individuals had a missing value for at least one candidate variable or the outcome, and to avoid the need for possibly incorrect assumptions, a complete-case analysis was used for the external validation. The internal and external AUROC, along with 95% confidence interval (CI), was calculated for the two HbA1c RATs developed in this chapter as well as for the LSA risk score. The internal and external Brier scores were calculated for the three RATs, along with the outcome index variance for each dataset.

Cut-points were chosen to create four risk groupings for both of the HbA1c RATs, these cut-points were based on the sensitivity and specificity of the RATs for the outcome in the internal data. This allowed the Net Reclassification Improvement (NRI) of the potential new RATs' groups compared to the existing LSA risk groups to be calculated in the external data, along with 95% CIs. Additionally, the NRI of the electronic RAT's risk groups compared to the pen and paper RAT's risk groups was calculated, along with 95% CI. The NRI shows the difference implementing one RAT in place of another would make to the risk group, and thus the advice given to individuals. It should be noted that the middle cut-point is the most important as this determines whether individuals receive a blood screening test or not.

6.4 Results

There were 6,390 individuals aged 40- 75 years old in the ADDITION-Leicester dataset, Table 6.1 displays the amount of missing data for each candidate variable and for the outcome variable. The 85 individuals without the outcome variable were excluded after the multiple imputation had been carried out, leaving 6,305 individuals to develop the two RATs with. Table 6.1 shows the imputed data is similar to the observed data, however there are slightly higher levels of several of the risk factors in the imputed data than the observed data.

Table 6.1 Summary statistics of outcome and candidate variables in ADDITION-Leicester dataset

Variable	Observed data only	Observed and imputed data	Number of Missing values (%)
NDH/undiagnosed T2DM by HbA1c (%)	23.2	N/A	85 (1.3)
Age, years	57.3 (9.60)	57.4 (9.59)	1 (0.0)
Sex, Male (%)	49.9	47.7	0 (0.0)
Ethnicity, White European (%)	75.8	76.1	103 (1.6)
BMI (kg/m^2)	28.1 (4.99)	28.1 (4.97)	221 (3.5)
Waist (cm)	94.2 (13.1)	94.2 (13.1)	225 (3.5)
Current smoker (%)	14.5	14.5	237 (3.7)
*Used high blood pressure drugs (%)	23.4	24.8	1,232 (19.3)
Previous stroke (%)	2.1	2.8	1,646 (25.8)
History of high cholesterol (%)	17.4	19.0	1,530 (23.9)
*History of high blood pressure (%)	27.8	29.4	1,438 (22.5)
History of Angina (%)	4.8	5.7	1,657 (25.9)
1st Degree Relative with diabetes (%)	25.2	26.0	1,204 (18.8)
Females with history of gestational diabetes (%)	1.3	1.3	0 (0)
Females with PCOS (%)	0.5	0.5	0 (0)
On steroids (%)	5.1	5.1	0 (0)

Values given as: mean (sd), unless stated

*These two high blood pressure variables were combined after imputation to make the combined hypertension candidate variable: history of high blood pressure or antihypertensive use.

6.4.1 Electronic risk assessment tool development

Table 6.2 displays the logistic regression model selected by backward elimination for the outcome of HbA1c ≥6.0%.

Table 6.2 Logistic regression model for outcome of HbA1c ≥6.0% selected using backward elimination starting with all candidate variable (n=6,305)

Variable		Coefficient	95% CI	P Value
Age (years)		0.0477	0.040, 0.055	<0.001
Angina	No	Reference group		
	Yes	0.510	0.22, 0.80	0.001
BMI (kg/m^2)		0.0564	0.037, 0.076	<0.001
Current smoker	No	Reference group		
	Yes	0.458	0.28, 0.64	<0.001
Ethnicity	White	Reference group		
	Other	1.28	1.1, 1.4	<0.001
First degree family history of	No	Reference group		
T2DM	Yes	0.344	0.19, 0.50	<0.001
Waist circumference (cm)		0.0398	0.0050, 0.021	0.001

Table 6.3, displays the logistic regression model with linear terms only chosen once the previous evidence and health message of variables had been taken into account, in addition to the automated procedure. Angina, which was selected by the automatic procedure, was replaced by hypertension in the model. Hypertension has frequently been included in RATs for binary glucose outcomes, whereas angina very rarely has (55-59,99). The internal AUROC only dropped slightly as a result and thus hypertension replaced angina due to the strong support for the inclusion of the hypertension variable from previous evidence.

Table 6.3 Logistic regression model for outcome of HbA1c ≥6.0% selected after considering the previous evidence and health message (n=6,305)

Variable		Coefficient	95% CI	P Value
Age (years)		0.0483	0.041, 0.056	<0.001
BMI (kg/m^2)		0.0540	0.034, 0.074	<0.001
Current smoker	No	Reference group		
	Yes	0.463	0.28, 0.65	<0.001
Ethnicity	White	Reference group		
	Other	1.28	1.1, 1.4	<0.001
First degree family history of	No	Reference group		
T2DM	Yes	0.35	0.20, 0.50	<0.001
Hypertension (history of high	No	Reference group		
blood pressure or	Yes	0.169	0.016, 0.32	0.030
antihypertensive use)				
Waist circumference (cm)		0.0133	0.0055, 0.021	0.001

Table 6.4 displays the logistic regression model selected for the electronic RAT. This model used a forward stepwise procedure to add interaction and quadratic terms to the linear terms selected for inclusion in the model displayed in Table 6.3. A quadratic term for age has been added, as well as an interaction between current smoker and ethnicity. The electronic RAT would give individuals completing it their probability of having HbA1c ≥6.0% based on their values of the seven variables included in this logistic regression model.

Table 6.4 Logistic regression model of electronic risk assessment tool for outcome of HbA1c ≥6.0% (n=6,305)

Terms		**Coefficient**	**95% CI**	**P Value**
Constant		-10.3	-12.7,-7.8	<0.001
Age (years)		0.147	0.065, 0.23	<0.001
Age*Age (years^2)		-0.000850	-0.0016, -0.00014	0.018
BMI (kg/m^2)		0.0523	0.033, 0.072	<0.001
Current smoker	No	Reference group		
	Yes	0.590	0.39, 0.79	<0.001
Ethnicity	White	Reference group		
	Other	1.35	1.2, 1.5	<0.001
Current Smoker*Ethnicity		Reference group		
	Smoker & Other	-0.598	-1.0,-0.16	0.008
First degree family history of T2DM	No	Reference group		
	Yes	0.357	0.20, 0.51	<0.001
Hypertension (history of high blood pressure or antihypertensive use)	No	Reference group		
	Yes	0.170	0.017, 0.32	0.029
Waist circumference (cm)		0.0139	0.0065, 0.23	<0.001

Table 6.5 shows the predictive diagnostics of the cut-points chosen to create the electronic score risk groups (low, moderate, high and very high) in the internal dataset. The cut-points were chosen to give a sensible balance of sensitivity and specificity given the risk groups they are thresholds for. The predictive diagnostics of these cut-points for the HbA1c based outcome are similar to those observed for their LSA counterparts for the original outcome the LSA was developed using (51).The middle cut-point, ≥0.2 in this case, is the most important as it is the cut-point which would be used for the decision of whether to offer a blood screening test or not.

Table 6.5 Proportion high risk, sensitivity, specificity, PPV and NPV of cut-points selected for electronic risk assessment tool in the internal dataset (n=6,305)

Cut-points	Proportion high risk	Sensitivity	Specificity	PPV	NPV
≥0.1	85.2	95.9	18.0	26.1	93.5
	(84.3, 86.1)	(94.8, 96.9)	(16.9, 19.1)	(24.9, 27.3)	(91.9, 95.2)
≥0.2	51.2	73.6	55.5	33.4	87.4
	(49.9, 52.5)	(71.2, 76.0)	(54.1, 57.0)	(31.7, 35.0)	(86.2, 88.6)
≥0.4	12.0	25.3	92.1	49.1	80.3
	(11.1, 12.8)	(23.0, 27.6)	(91.3, 92.8)	(45.4, 52.7)	(79.2, 81.3)

Data given as: % (95% CI)

6.4.2 Pen and paper risk assessment tool development

Table 6.6 displays the logistic regression model yielded from using the same variables selected by backward elimination earlier but with the continuous variables grouped.

Table 6.6 Logistic regression model for outcome of HbA1c ≥6.0% using the same variables as in Table 6.1 but with continuous variables grouped (n=6,305)

Variable	Grouping	Coefficient	95% CI	P Value
Age (years)	40-49	Reference group		
	50-59	0.514	0.33, 0.70	<0.001
	60-69	1.01	0.81, 1.2	<0.001
	70-75	1.22	0.99, 1.4	<0.001
Angina	No	Reference group		
	Yes	0.550	0.26, 0.84	<0.001
BMI (kg/m^2)	<25	Reference group		
	25-29	0.194	0.0038, 0.38	0.046
	30-34	0.505	0.26, 0.74	<0.001
	≥35	0.765	0.46, 1.1	<0.001
Current smoker	No	Reference group		<0.001
	Yes	0.408	0.23, 0.59	<0.001
Ethnicity	White	Reference group		
	Other	1.25	1.1, 1.4	<0.001
First degree family history of T2DM	No	Reference group		
	Yes	0.339	0.18, 0.49	<0.001
Waist circumference (cm)	<90	Reference group		
	90-99	0.239	0.056, 0.42	0.010
	100-109	0.278	0.054, 0.50	0.015
	>109	0.733	0.45, 1.0	<0.001

Table 6.7 displays the logistic regression model and scoring system of the pen and paper RAT selected. The variables included differed slightly to those selected automatically, with angina being replaced by hypertension, as was the case for the electronic RAT. The internal AUROC of the risk score only dropped very slightly as a result of replacing angina with hypertension and thus due to the existing evidence supporting hypertension as a risk factor it replaced angina in the pen and paper RAT.

Table 6.7 Logistic regression model and associated scoring system for outcome of HbA1c ≥6.0% grouped continuous variables and taking previous evidence and health message into account when selecting variables (n=6,305)

Variable Grouping		**Coefficient**	**95% CI**	**P Value**	**Scoring**
Age (years)	40-49	Reference group			0
	50-59	0.503	0.31, 0.69	<0.001	5
	60-69	1.00	0.81, 1.2	<0.001	10
	70-75	1.23	1.0, 1.4	<0.001	12
BMI (kg/m^2)	<25	Reference group			
	25-29	0.181	-0.0094, 0.37	0.062	2
	30-34	0.484	0.24, 0.72	<0.001	5
	≥35	0.720	0.41, 1.0	<0.001	7
Current smoker	No	Reference group			
	Yes	0.416	0.23, 0.60	<0.001	4
Ethnicity	White	Reference group			
	Other	1.24	1.1, 1.4	<0.001	12
First degree family history of T2DM	No	Reference group			
	Yes	0.345	0.19, 0.50	<0.001	3
Hypertension (history of high blood pressure or antihypertensive use)	No	Reference group			
	Yes	0.195	0.043, 0.35	0.012	2
Waist circumference (cm)	<90	Reference group			
	90-99	0.239	0.057, 0.42	0.010	2
	100-109	0.284	0.060, 0.51	0.013	3
	>109	0.744	0.46, 1.0	<0.001	7

Table 6.8 displays the predictive diagnostics of the cut-points chosen to create the pen and paper RAT's risk groups (low, moderate, high and very high) in the internal dataset. The cut-points were chosen to give a sensible balance of sensitivity and specificity given the risk groups they are thresholds for, although these happened to be the same as the LSA cut-points currently used in practice. The predictive diagnostics are comparable to those yielded by their LSA counterparts for the original outcome the LSA was developed using (51).

Table 6.8 Proportion high risk, sensitivity, specificity, PPV and NPV of cut-points selected for the pen and paper risk assessment tool in internal dataset (n=6,305)

Cut-points	Proportion high risk	Sensitivity	Specificity	PPV	NPV
≥7	10.3	97.4	12.6	25.2	94.2
	(9.5, 11.0)	(96.6, 98.3)	(11.7, 13.6)	(24.1, 26.3)	(92.3, 96.1)
≥16	52.1	74.2	54.6	33.1	87.5
	(50.8, 53.3)	(71.8, 76.6)	(53.2, 56.0)	(31.5, 34.7)	(86.3, 88.7)
≥25	13.0	26.6	91.2	47.7	80.4
	(12.1, 13.8)	(24.2, 29.0)	(90.3, 92.0)	(44.1, 51.2)	(79.4, 81.5)

Data given as: % (95% CI)

6.4.3 Comparison of HbA1c risk assessment tools and LSA risk score

Table 6.9 displays the internal and external AUROCs of the two RATs developed in this chapter as well as those of the LSA risk score for the outcome of HbA1c ≥6.0%. The AUROCs, show good levels of discrimination by all three RATs for the outcome of abnormal glucose. The external AUROC of the electronic RAT is slightly but significantly higher than both the LSA's AUROC and the pen and paper RAT's AUROC. This indicates the electronic RAT discriminates between individuals with and without abnormal HbA1c marginally better than the other two RATs. The external AUROC of the pen and paper RAT is not significantly different to the LSA's (p=0.39). The internal and external Brier scores of the three RATs associated probabilities are shown in Table 6.9. All are below the outcome index variance of each dataset, showing their predictions outperform assigning each individual a prediction of the prevalence of the dataset. The LSA's Brier scores are the highest which is to be expected as it was not developed for the outcome being assessed.

Table 6.9 Internal (n=6,305) and external (n=3,165) AUROCs and Brier scores of the LSA risk score, the pen and paper risk assessment tool and the electronic risk assessment tool

Risk assessment tool	**Internal AUROC** (95% confidence Interval)	**External AUROC** (95% confidence Interval)	**Internal Brier score** (outcome index variance)	**External Brier score** (outcome index variance)
LSA	0.675 (0.66, 0.69)	0.671 (0.65, 0.69)	0.167 (0.178)	0.196 (0.207)
Electronic	0.708 (0.69, 0.72)	0.690 (0.67, 0.71)	0.161 (0.178)	0.191 (0.207)
Pen and paper	0.702 (0.69, 0.72)	0.677 (0.66, 0.70)	0.162 (0.178)	0.192 (0.207)

Figures 6.2-6.4 show that the RATs have similar calibration in the external dataset. They all underestimate the risk of the outcome slightly since the prevalence of the outcome is higher in the external dataset than the internal dataset.

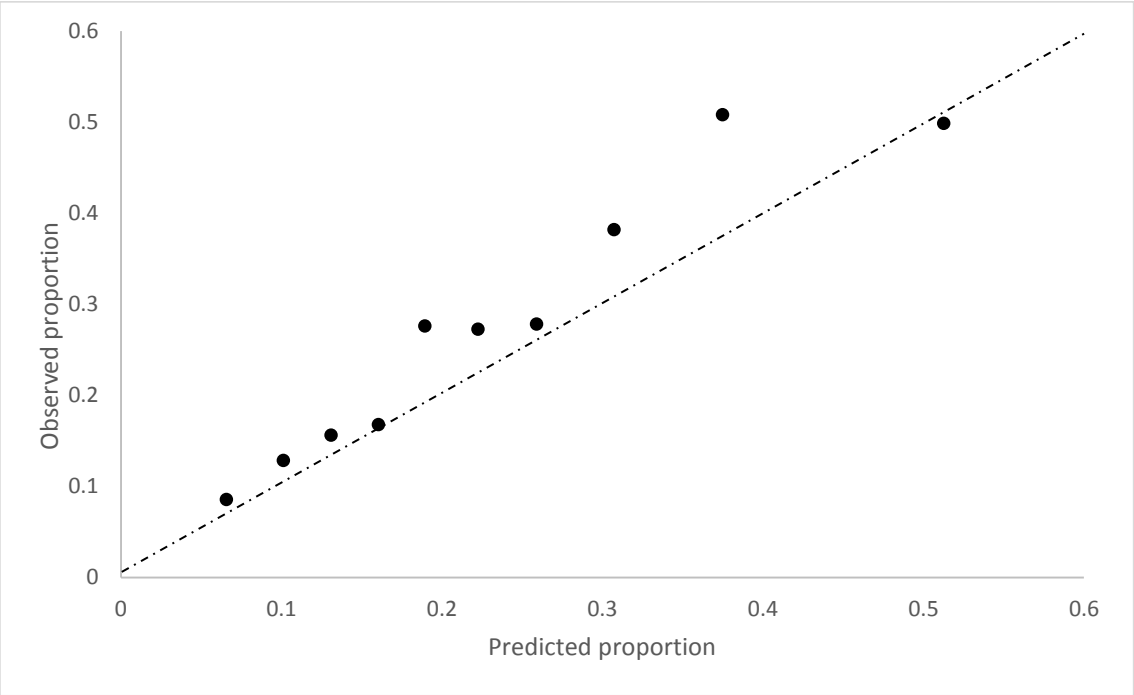


Figure 6.2 Predicted against observed risk of outcome by decile for LSA risk assessment tool in external dataset

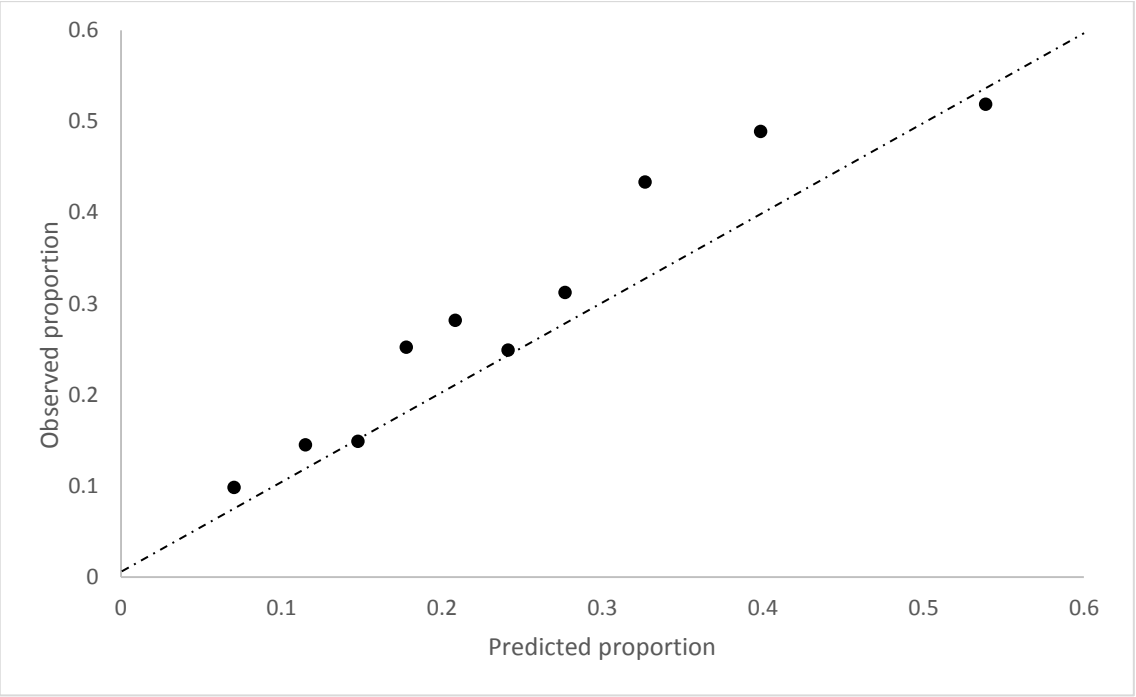


Figure 6.3 Predicted against observed risk of outcome by decile for electronic risk assessment tool in external dataset

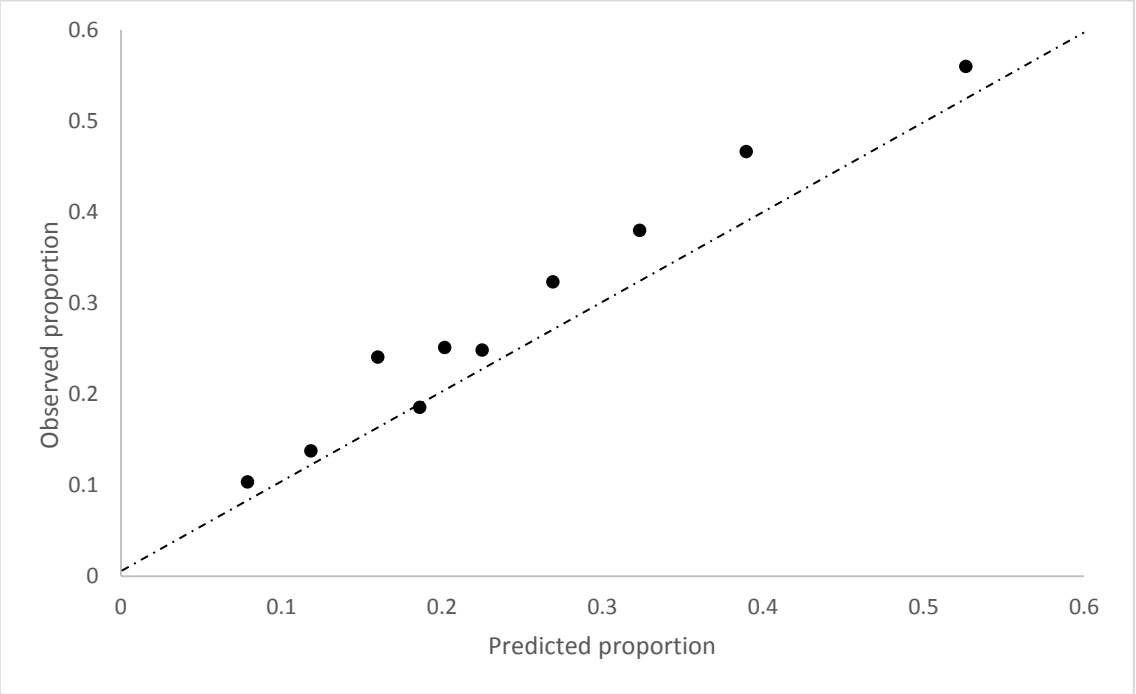


Figure 6.4 Predicted against observed risk of outcome by decile for pen and paper risk assessment tool in external dataset

Table 6.10 compares the risk groups given to individuals in the external dataset by the pen and paper RAT to those given by the LSA risk score. The NRI of -3.24% indicates the pen and paper risk groups on average were slightly worse at categorising individuals than the LSA risk groups, although it was not significant at the 5% level.

Table 6.10 Risk Classification of individuals in the external dataset under pen and paper risk assessment tool groups compared to LSA risk groups split by HbA1c status (n=3,165)

Risk classification using LSA score	**Risk classification using new grouped score**				**Reclassified**		
	Low	Moderate	High	Very High	Higher risk	Lower risk	Net correctly reclassified
Cases (*n*=927)							
Low	14	6	0	0			
Moderate	6	94	44	0	88	234	-15.7%
High	0	95	278	38			
Very High	0	0	133	219			
Non cases (*n*=2,238)							
Low	117	50	0	0			
Moderate	39	527	165	0	258	538	12.5%
High	0	281	613	43			
Very High	0	2	216	185			
NRI (95% CI)					-3.24% (-7.64, 1.38)		

Table 6.11 compares the risk groups given to individuals in the external dataset by the electronic RAT to those given by the LSA risk score. The NRI of 2.09% indicates the electronic risk groups on average were marginally better at categorising individuals than the LSA risk groups, although it was not significant at the 5% level.

Table 6.11 Risk Classification of individuals in the external dataset under electronic risk assessment tool groups compared to LSA risk groups split by HbA1c status (n=3,165)

Risk classification using LSA score	**Risk classification using new continuous score**				**Reclassified**		
	Low	Moderate	High	Very High	Higher risk	Lower risk	Net correctly reclassified
Cases (*n*=927)							
Low	13	7	0	0			
Moderate	20	81	43	0	91	280	-20.4%
High	4	100	266	41			
Very High	0	2	154	196			
Non cases (*n*=2,238)							
Low	118	55	0	0			
Moderate	170	437	118	0	232	735	22.5%
High	16	317	545	59			
Very High	0	4	228	171			
NRI (95% CI)					2.09% (-2.42, 6.86)		

Table 6.12 compares the risk groups given to individuals in the external dataset by the electronic RAT to those given by the pen and paper RAT. The NRI of 4.60% indicates the electronic risk groups on average were better at categorising individuals than the pen and paper risk groups, this was significant at the 5% level.

Table 6.12 Risk Classification of individuals in the external dataset under electronic risk assessment tool groups compared to pen and paper risk assessment tool groups split by HbA1c status (n=3,165)

Risk classification using new grouped score	**Risk classification using new continuous score**				**Reclassified**		
	Low	Moderate	High	Very High	Higher risk	Lower risk	Net correctly reclassified
Cases (*n*=927)							
Low	16	4	0	0			
Moderate	14	147	34	0	67	108	-4.42%
High	7	38	381	29			
Very High	0	1	48	208			
Non cases (*n*=2,238)							
Low	139	17	0	0			
Moderate	121	635	104	0	173	375	9.03%
High	44	160	738	52			
Very High	0	1	49	178			
NRI (95% CI)					4.60% (1.08, 8.19)		

6.5 Discussion

The LSA risk score is a widely used self-assessment RAT which identifies individuals at a high risk of currently having NDH or undiagnosed T2DM and therefore should be screened using a blood test. Since the LSA was developed and validated for abnormal glucose defined by an OGTT, this chapter considered whether it needs to be replaced with a RAT developed for the outcome of abnormal HbA1c in light of the increased use of HbA1c as the diagnostic for diabetes status. Two self-assessment RATs for the outcome of HbA1c ≥6.0% were developed, a pen and paper RAT and an electronic RAT. The comparison carried out in external data did not find either RAT outperformed the existing LSA risk score considerably, additionally neither of the RAT's risk groups outperform the established LSA risk groups; therefore neither are recommended to replace the LSA risk score in practice. Although not selected in either the pen and paper RAT or the electronic RAT, Angina was included in the model chosen by backward elimination showing it is a potential risk factor for abnormal HbA1c.

6.5.1 Comparison of risk assessment tools and LSA risk score

The pen and paper RAT had similar discrimination to the existing LSA risk score in the external dataset, with a test of difference in the AUROCs being non-significant (p=0.39). The NRI comparing the risk groups indicated the pen and paper risk groups on average reclassified more individuals incorrectly than correctly compared to the LSA risk groups, although this too was non-significant. Based on the AUROCs and the NRI in the external dataset the pen and paper RAT developed in this chapter should not replace the existing LSA risk score.

The electronic RAT had slightly better discrimination and calibration than the LSA risk score in the external dataset. Additionally, the NRI comparing the risk groups showed the electronic risk groups on average reclassified more individuals correctly than incorrectly compared to the LSA risk groups, although this was not significant and was only 2.1%. Since the electronic RAT is more difficult to calculate, fewer individuals completing it would understand the importance of the different risk factors on their overall risk than would when completing the LSA risk score. Due to this loss of an educational message

about risk factors and there not being a significant difference in the performance of the risk groups of the two RATs, which is likely to be an individuals' main understanding of their risk upon completing a RAT, the LSA risk score should not be replaced by the electronic RAT.

The electronic RAT had moderately better discrimination than the pen and paper RAT. The electronic risk groups outperformed the pen and paper risk groups with the NRI being significantly positive when considering reclassifying individuals from the pen and paper risk groups to the electronic risk groups. With both the AUROC and NRI being significant but modestly in favour of the electronic RAT, if both of the RATs had been suitable to replace the LSA risk score, careful consideration would be required to decide which one should be selected; as the pen and paper RAT has the advantage of disseminating an educational message more clearly than the electronic RAT. The electronic RAT outperforming the pen and paper RAT is to be expected, with a study of the impact of categorising risk factors in RATs for longitudinal outcomes showing it reduces the performance (187). Furthermore, the relatively small decreases in performance metrics are likely due to four categories being used for each continuous variable, the comparison study found that using more categories lessens the negative impact on performance.

6.5.2 Strengths and weaknesses of analysis

One advantage of this comparison was the use of the same dataset to develop the RATs as was used to develop the LSA risk score. Although not all the same candidate variables were considered, as the LSA considered a few variables which were not included in the external dataset used in this comparison. Notably, fruit and diet intake and physical activity were not considered for the new RATs developed. Chapter 3 found diet factors and physical activity were often included when considered for RATs developed using logistic regression, although the LSA did not include these variables.

A strength of the RATs developed was the use of multiple imputation of missing candidate variables in the internal dataset, in which around a third of individuals had at least one candidate variable missing. Multiple imputation is recommended as it reduces the loss of information and reduces the bias

caused by the missing data (130,136,140). 50 imputations were used which simulation studies suggest is more than sufficient (145). Additionally as is best practice, all available variables believed to explain the missing variable being modelled were used in the multiple imputation even if they were not in the subsequent analysis (140). Another strength of the RATs developed was that variables were selected using statistical methods alongside previous evidence and the health message given by implementing the RAT in practice (66,69).

Importantly, the comparison considered benefits of implementing the RATs in practice. In addition to the discrimination of the RATs for the outcome of interest, the health messages the RATs would provide to individuals completing them and the performance of the RATs' risk groups were compared. The performance of the RATs' risk groups were compared using an unweighted NRI, this has the disadvantage of only considering whether a reclassification is in the right direction and not the importance of moving from the first category to the second (190,191). This means moving from low to moderate is equally important to the measure as moving from moderate to high. Although this is not ideal for the comparison using NRI is important as the advice an individual receives when completing a self-assessment depends on the risk group they receive and thus the main impact in practice is determined by the risk group given.

6.6 Conclusion and implications

The pen and paper RAT's performance was very similar to the LSA score's for identifying individuals with abnormal HbA1c; while the electronic RAT which has practical disadvantages saw only a modest improvement in performance, which was not significant in terms of risk groups. Therefore, neither RAT should replace the LSA score even if the second stage of screening is HbA1c testing, as the LSA was not meaningfully outperformed by either RAT. Most RATs developed for the outcome of current NDH or undiagnosed diabetes define NDH by OGTT (99). These findings give some reassurance that these RATs may perform similarly well in identifying HbA1c NDH or undiagnosed diabetes as ones developed specifically for this outcome. However as RATs performance is population specific the performance of other RATs for identifying abnormal HbA1c should be validated externally before using them to identify individuals to be given HbA1c tests. Chapter 8 will add to the evidence of whether the LSA risk score is appropriate for identifying individuals with abnormal HbA1c, by assessing the performance of the LSA for identifying abnormal HbA1c in a dataset of individuals from across England.

Chapter 7: Establishing risk groups for the Leicester Practice Risk Score

7.1 Chapter Outline

This chapter establishes risk groupings for the Leicester Practice Risk Score (LPRS) in order to enable consistent advice to be given across different general practices when utilising the risk assessment tool (RAT). Four risk groupings (low, moderate, high and very high) are created and validated for the outcome of currently having an abnormal glucose level. In addition, as risk groupings have already been established for the Leicester Self-Assessment (LSA) risk score, high agreement between the risk groupings developed for the LPRS in this chapter and those already established for the LSA risk score was aimed for.

The work in this chapter has been:

- Orally presented:
  Barber SR, Davies MJ, Khunti K, Gray LJ. 'Establishing risk groupings for the Leicester Practice Risk Score'. At: The Society for Academic Primary Care Trent Regional Spring Meeting. 15th March 2016. (P9)

7.2 Introduction

As detailed in Chapter 1, the LPRS was developed in order for general practices to utilise data routinely stored in primary care to identity individuals most likely to have non-diabetic hyperglycaemia (NDH) or undiagnosed Type 2 diabetes mellitus (T2DM), and therefore requiring a blood test. The LPRS was developed using a logistic regression model with six variables commonly stored in primary care databases. Although not original designed to calculate a probability for the outcome, the coefficients given for the LPRS can be used to calculate the probability of an individual currently having abnormal glucose. The equation for calculating the LPRS as a probability is detailed in (7.1) with the coefficients rounded to 3 significant figures.

$$\text{LPRS} = \frac{1}{1 + e^{-\left(-6.78 + (0.0401 * age) + (0.0821 * BMI) + (0.184 * sex) + (0.757 * BME) + (0.550 * HBP) + (0.477 * FHD)\right)}} \quad (7.1)$$

Where

Sex=0 if female and 1 if male

BME=0 if White European and 1 if other ethnicity

HBP=0 if not prescribed anti-hypertensives and 1 if prescribed anti-hypertensives

FHD=0 if no 1st degree family history of diabetes or 1 if family history of diabetes

The LPRS has been externally validated with good discrimination found and its implementation in two prevention trials has shown its use greatly reduces the cost of detecting individuals who currently have abnormal glucose levels compared to population level screening (45,52). The LPRS was created to allow primary care practices, wishing to perform a mass invitation to screening, to rank individuals by their risk of abnormal glucose (45). To date, it has only been available as an add-on to software, meaning practices carrying out targeted screening programmes can use it to rank individuals and chose a threshold depending on the resources they have available for this screening programme. This has limited the use of the LPRS in practice. However, the LPRS is currently being incorporated by two providers into their database systems which are used by many general practices, namely Vision and SystmOne. Supplying a risk group for each individual alongside their LPRS would empower general practitioners to be able to easily use the LPRS for opportunistic screening in consultations which are already taking place, in addition to any target screening

invitations their general practice as a whole decides to carry out. The National Health Service (NHS) Diabetes Prevention Programme (DPP) highlights that such opportunistic screening should be carried out to help identify individuals with NDH or undiagnosed T2DM (36). Furthermore, providing risk groups ensures consistent advice and screening decisions will be given across different general practices when utilising the RAT; as the advice and screening decision individuals receive currently depends on their practice. Therefore this chapter aimed to propose four risk groupings (low, moderate, high and very high) of the LPRS to be used universally. Individuals within the low or moderate risk groupings should be considered to have screened negative using the LPRS and in practice should be given advice on how to stay healthy or lower their risk score. While individuals within the high or very high risk groupings should be considered to have screened positive using the LPRS and so in addition to being informed how to lower their risk they should be offered a blood screening test, glycated haemoglobin A1c (HbA1c) or fasting plasma glucose (FPG), as specified by the National Institute for Health and Care Excellence (NICE) (28).

The LSA risk score is a widely available and used simple risk score that individuals can calculate themselves; it has been detailed in Chapter 1 (46). It uses seven risk factors, six of which are the same or very similar to the risk factors used to calculate the LPRS. Since the risk scores are heavily related ideally they should communicate the same or at least a similar message to individuals about their risk. As risk groupings have already been established for the LSA risk score, this work aimed to achieve high agreement between the risk groupings advocated for the LPRS in this chapter and those already established for the LSA risk score. Furthermore, high agreement in the screening decisions suggested in this chapter and those given by using the LSA with recommended cut-point were aimed for.

7.3 Methods

The LPRS was calculated as a probability using equation (7.1). The ADDITION-Leicester dataset was used to develop two sets of risk groupings which were considered for recommending for use across general practices. The most-up-to-date version of the ADDITION-Leicester dataset contained 6,075 individuals aged 40- 75 years old with complete data for all the risk factors for the LPRS and LSA risk score as well as the outcome. The most up-to-date version of the Screening those at risk (STAR) dataset contained 2,872 with complete information required for the work in this chapter. The ADDITION-Leicester and STAR datasets are detailed in Chapter 2.

The binary outcome of abnormal glucose used for the analysis in this chapter was HbA1c ≥6.0%. This is due to HbA1c being more commonly used by general practices as the test to define glucose status, and the NICE guidelines specifying this as the cut-point for defining abnormal glucose which requires an intervention to be offered (28). A sensitivity analysis was carried out with the outcome being defined using FPG, abnormal glucose was defined as FPG ≥5.5mmol/l as stipulated in the two-stage screening programme advocated in the NICE guidelines (28).

The first set of risk groups, which will be referred to as the Initial risk groups, were chosen so that the probabilities of their cut-points closely matched the probabilities associated with the LSA risk groups' cut-points. Additionally, they were chosen so that individuals with the same rounded percentage probability of the outcome were grouped together, i.e. they were of the form 0. _ _ 5. This means individuals could be given their percentage to the nearest whole number if they ask for it in a consultation without needing to worry that they may speak to another individual who has been given the same percentage but different risk grouping leading to confusion.

The second set of risk groups considered for the LPRS, which will be referred to as the Simplified risk groups, were chosen based on the predictive diagnostics in the internal dataset. Also the risk groups were chosen to be easy to remember, as the widely used body mass index (BMI) groups are, for example overweight has cut-points of ≥25 and ≤30. Again, the groups were chosen so

individuals with the same rounded percentage probability of the outcome were grouped together.

Sensitivity, specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) of the cut-points for both sets of risk groups in the ADDITION-Leicester dataset were calculated, along with their 95% confidence intervals (CI), to assess their internal performance. The level of agreement, expected level of agreement and Cohen's kappa coefficient (described in subsection 7.3.1) with the LSA risk groups were calculated for both the Initial risk groups and the Simplified risk groups in the ADDITION-Leicester dataset. Additionally, the level of agreement, expected level of agreement and Cohen's kappa coefficient with the LSA screening decisions were calculated for both the screening decisions of the Initial risk groups and the screening decisions of the Simplified risk groups. 95% CIs were also calculated for the kappa coefficients.

The STAR dataset was used to assess the external performance of the cut-points of both the Initial risk groups and the Simplified risk groups. The sensitivity, specificity, PPV and NPV of each cut-point of the Initial and Simplified risk groups were calculated, along with their 95% CIs. The level of agreement, expected level of agreement and Cohen's kappa coefficient of the LSA risk groups with both the Initial risk groups and the Simplified risk groups in the STAR dataset were calculated. Finally, the level of agreement, expected level of agreement and Cohen's kappa coefficient of the LSA screening decisions with both the Initial risk groups' screening decisions and the Simplified risk groups' screening decisions were evaluated. 95% CIs were also calculated for the kappa coefficients.

7.3.1 Cohen's kappa coefficient

Cohen's Kappa coefficient , κ, is a statistical metric for inter-rater reliability; it details the levels of agreement in classifications given by two different raters (192). In the context of this chapter, κ measures how often categorisations of individuals' risk determined by two different risk scores' agree.

Equation (7.2) shows that κ reports how the observed level of agreement , P_O, compares to the expected level of agreement by chance, P_E (192). As (7.2) states, P_O is calculated by summing the number of individuals given a matching

categorisation for each of the categories, M_i, and dividing by the total number of individuals rated, T. While P_E is calculated by summing the products of the proportion of individuals classified into a category by rater A, $\frac{A_i}{T}$, with the proportion of individuals classified into that category by rater B, $\frac{B_i}{T}$. Where A_i and B_i are the number of individuals placed into category i by rater A and rater B respectively.

$$\kappa = \frac{P_O - P_E}{1 - P_E} \tag{7.2}$$

$$\text{where } P_E = \sum_{i=1}^{n} \frac{A_i}{T}\frac{B_i}{T} \quad \& \quad P_O = \frac{\sum_{i=1}^{n} M_i}{T}$$

A κ of one is yielded when there is perfect agreement between the two groups of categorisations; while a κ of below zero shows the agreement is worse than expected by chance. Table 7.1 displays the strength of agreement different κ values indicate according to Landis and Koch (193).

Table 7.1 Strength of agreement shown by different values of Kappa according to Landis and Koch

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

7.4 Results

In the internal dataset 22.6% of the 6,075 individuals had HbA1c ≥6.0%. While in the external dataset 29.0% of the 2,872 individuals had the outcome.

7.4.1 Initial risk groups

Table 7.2 details the cut-points for the Initial risk groups considered for the LPRS, along with the probability of having the outcome associated with the LSA risk groups' cut-points which they were chosen to closely match. These cut-points lead to the following LPRS risk groups when displaying the probabilities as percentages:

- Low risk: 0- 7%
- Moderate risk: 8-16%
- High risk: 17- 32%
- Very high risk: 33% or more

Table 7.2 Cut-points for Initial LPRS risk groups which closely match the probability associated with LSA risk groups' cut-points

LSA cut-points	Probability associated with LSA score	Initial cut-point proposed for LPRS
≥7	0.0722	≥0.075
≥16	0.1608	≥0.165
≥25	0.3204	≥0.325

The sensitivity, specificity, PPV and NPV of the Initial risk groups' cut-points in the internal and external datasets are displayed in Table 7.3. The sensitivities and specificities of the cut-points were comparable in the internal and external data. The PPVs were higher in the external dataset compared to the internal dataset; while the NPVs decreased in the external dataset. The cut-point of ≥0.075 had extremely high sensitivity and very low specificity in both the internal and external data. The cut-point of ≥0.165, which is the most important as it is the cut-point for the decision of whether individuals are offered a blood screening test, had good sensitivity but modest specificity in both datasets. The cut-point of ≥0.325 had good specificity and low sensitivity in both datasets.

Similar values of sensitivity, specificity, PPV and NPV were found in the sensitivity analysis, were the outcome was FPG ≥5.5mmol/l, displayed in the Appendix B.

Table 7.3 Sensitivity, specificity, PPV and NPV of the Initial risk groups' cut-points in the internal and external datasets

Cut-point	ADDITION-Leicester				STAR			
	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
≥0.075	97.5	12.3	24.5	94.4	97.5	9.8	30.7	90.5
	(96.6, 98.3)	(11.3, 13.2)	(23.4, 25.7)	(92.3, 96.1)	(96.2, 98.4)	(8.6, 11.1)	(28.9, 32.5)	(85.8, 94.0)
≥0.165	73.7	53.3	31.6	87.4	72.9	51.7	38.2	82.3
	(71.3, 76.0)	(51.8, 54.7)	(30.0, 33.2)	(86.1, 88.6)	(69.7, 75.9)	(49.5, 53.9)	(35.8, 40.6)	(80.1, 84.4)
≥0.325	27.9	89.9	44.6	81.0	21.7	90.3	47.9	73.8
	(25.5, 30.3)	(89.0, 90.7)	(41.2, 48.0)	(79.9, 82.1)	(18.9, 24.7)	(89.0, 91.6)	(42.7, 53.1)	(72.0, 75.5)

Values given as % (95% CI).

Table 7.4 and Table 7.5 display the proportion of individuals that were categorised into each combination of the LSA risk groups and Initial LPRS risk groups in the internal and external datasets respectively. A higher proportion of individuals were classified by the LSA risk groups as high and very high in the external dataset compared to the internal dataset. The internal and external datasets had a similar proportion of individuals in each of the Initial LPRS risk group. The Initial LPRS risk groups were the same as the LSA risk groups for 65.6% of individuals in the external dataset, a slight decrease from the internal dataset were this was the case for 68.6% of individuals. Table 7.4 indicates only 0.4% of individuals in the internal dataset had an Initial LPRS risk group that differ by two or more categories compared to their LSA risk group. While Table 7.5 displays only 0.3% of individuals in the external dataset had an Initial LPRS risk group that differ by two or more categories compared to their LSA risk group.

Table 7.4 Frequency of LSA risk groups compared to Initial LPRS risk groups in the internal dataset

		Initial LPRS risk group				
		Low	Moderate	High	Very High	Total
LSA risk group	Low	7.2	3.2	0.0	0.0	10.5
	Moderate	2.8	24.7	6.2	0.2	33.9
	High	0.0	9.1	26.5	3.8	39.4
	Very High	0.0	0.2	5.9	10.2	16.3
	Total	10.1	37.1	38.7	14.1	100

Values displayed as %.

Table 7.5 Frequency of LSA risk groups compared to Initial LPRS risk groups in the external dataset

		Initial LPRS risk group				
		Low	Moderate	High	Very High	Total
LSA risk group	Low	5.6	1.1	0.0	0.0	6.7
	Moderate	2.1	21.0	2.5	0.0	25.6
	High	0.0	14.4	27.2	1.3	42.9
	Very High	0.0	0.3	12.6	11.9	24.8
	Total	7.7	36.8	42.3	13.2	100

Values displayed as %.

Table 7.6 reports a kappa of 0.544 (95% CI: 0.528- 0.560) in the internal data and a kappa of 0.499 (95% CI: 0.476- 0.522) in the external data, this indicates moderate agreement in both datasets.

Table 7.6 Agreement of Initial LPRS risk groups with the LSA risk groups

Dataset	**Kappa** (95% CI)	**Agreement**	**Expected Agreement** (by chance)
Internal	0.544 (0.528, 0.560)	68.6%	31.2%
External	0.499 (0.476, 0.522)	65.6%	31.4%

Table 7.7 and Table 7.8 give the proportion of individuals that were classified as screening negative and positive by the Initial LPRS screening decision compared to the LSA screening decision in the internal and external datasets respectively. Lower proportions of individuals received a negative screening decision in the external dataset compared to the internal dataset using the LSA screening decision; however the proportions in the two datasets were similar using the Initial LPRS screening decision. The Initial LPRS screening decisions are the same as the LSA screening decisions for 82.7% of individuals in the external dataset, a slight decrease from the internal dataset where the two matched for 84.3% of individuals.

Table 7.7 Frequency of LSA screening decisions compared to Initial LPRS screening decisions in the internal dataset

		Initial LPRS screening decision		
		Negative	Positive	Total
LSA screening decision	Negative	37.9	6.4	44.3
	Positive	9.3	46.4	55.7
	Total	47.2	52.8	100

Values displayed as %.

Table 7.8 Frequency of LSA screening decisions compared to Initial LPRS screening decisions in the external dataset

		Initial LPRS screening decision		
		Negative	Positive	Total
LSA screening decision	Negative	29.8	2.5	32.3
	Positive	14.8	53.0	67.7
	Total	44.5	55.5	100

Values displayed as %.

Table 7.9 reports kappa coefficients of 0.685 (95% CI: 0.660- 0.710) in the internal data and 0.641 (95% CI: 0.605- 0.678) in the external data, this shows substantial agreement between the two sets of screening decisions.

Table 7.9 Agreement of Initial LPRS screening decisions with the LSA screening decisions

Dataset	**Kappa** (95% CI)	**Agreement**	**Expected Agreement** (by chance)
Internal	0.685 (0.660, 0.710)	84.3%	50.3%
External	0.641 (0.605 , 0.678)	82.7%	51.9%

7.4.2 Simplified risk groups

Table 7.10 displays the cut-points chosen for the Simplified risk groups considered for the LPRS. These cut-points lead to the following LPRS risk groups when displaying the probabilities as percentages:

- Low risk: 0- 10%
- Moderate risk: 11-15%
- High risk: 16- 30%
- Very high risk: 31% or more

The predictive diagnostics of Simplified risk groups' cut-points in the internal and external datasets are displayed in Table 7.10. The cut-points gave similar predictive diagnostics to those yielded by the Initial risk groups' cut-points. The predictive diagnostics yielded for each cut-points were sensible for the two risk groups they separated. The low/moderate cut-point being increased to ≥0.105 gave slightly lower sensitivities and increased, but still low, specificities. The moderate/high cut-point being decreased to ≥0.155 resulted in a small reduction in the high sensitivities seen and a slight increase in the moderate specificities. The high/very high cut-point being decreased to ≥0.305 increased the specificity and decreased sensitivity observed in both datasets.

Similar values of sensitivity, specificity, PPV and NPV were found in the sensitivity analysis, where the outcome was FPG ≥5.5mmol/l; displayed in Appendix B.

Table 7.10 Sensitivity, specificity, PPV and NPV of the Simplified risk groups' cut-points for HbA1c ≥6.0% in the internal and external datasets

Cut-point	ADDITION-Leicester				STAR			
	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
≥0.105	92.4 (90.9, 93.8)	26.0 (24.8, 27.3)	26.8 (25.5, 28.0)	92.2 (90.6, 93.6)	92.6 (90.6, 94.3)	23.3 (21.5, 25.2)	33.1 (31.2, 35.0)	88.5 (85.4, 91.0)
≥0.155	76.6 (74.2, 78.8)	49.4 (48.0, 50.9)	30.7 (29.1, 32.2)	87.8 (86.5, 89.0)	77.1 (74.1, 79.9)	47.3 (45.1, 49.5)	37.4 (35.2, 39.8)	83.5 (81.2, 85.6)
≥0.305	32.2 (29.8, 34.8)	87.6 (86.7, 88.6)	43.3 (40.2, 46.4)	81.6 (80.5, 82.6)	25.4 (22.5, 28.5)	88.1 (86.6, 89.5)	46.6 (41.9, 51.3)	74.3 (72.5, 76.0)

Values given as % (95% CI)

Table 7.11 reports the levels of agreement between the Simplified LPRS risk groups. The Kappa coefficients show the agreement is moderate when accounting for the expected agreement by chance. The Simplified LPRS risk groups agreement with the LSA risk groups was a little less than the agreement seen for the Initial LPRS risk groups in external dataset, 60.8% vs 65.6%.

Table 7.11 Agreement of Simplified LPRS risk groupings with the LSA risk groupings

Dataset	**Kappa** (95% CI)	**Agreement**	**Expected Agreement** (by chance)
Internal	0.478 (0.463, 0.493)	62.4%	28.0%
External	0.444 (0.423, 0.466)	60.8%	29.5%

The Simplified LPRS screening decisions matched the LSA screening decisions for 85.1% of individuals in the external dataset and 84.8% of individuals in the internal dataset. The kappa coefficients shown in Table 7.12 indicate substantial agreement between the two sets of screening decisions. The levels of agreement between the LSA and Simplified LPRS screening decisions in the external dataset were a little higher than those seen for the LSA and Initial LPRS screening decisions.

Table 7.12 Agreement of Simplified LPRS screening decisions with the LSA screening decisions

Dataset	**Kappa** (95% CI)	**Agreement**	**Expected Agreement** (by chance)
Internal	0.692 (0.666, 0.717)	84.8%	50.7%
External	0.680 (0.644, 0.716)	85.1%	53.5%

The two-way frequency tables comparing the Simplified LPRS risk groups to the LSA risk groups as well as the Simplified screening decision to the LSA screening decisions for the internal and external datasets are given in Appendix B.

7.5 Discussion

The use of a computer-based RAT by general practices for patients between 40 and 75 years old is recommended by NICE (28). The LPRS, unlike the LSA, is made up entirely of risk factors stored in primary care databases. The LPRS is being incorporated into some general practice database systems, namely Vision and SystmOne. Establishing risk groups for the LPRS enables general practitioners to use the LPRS easily for opportunistic screening in consultations. Furthermore, if adopted nationally the risk groups ensure the same advice and screening decision are given to an individual regardless of the general practice they are registered to, whether screening be opportunistic or using strategic invitations. Ideally the LPRS risk groups should be consistent with the LSA risk groups which are widely available for individuals to calculate themselves.

7.5.1 Performance of proposed risk groups and their agreement with LSA risk groups

Both the Initial and Simplified risk groups for the LPRS perform acceptably in terms of statistical performance. With the cut-points of both producing acceptable sensitivities and specificities for the two risk groups they separated. The Simplified risk groups' lowest cut-point had considerably higher specificity and slightly higher PPV than the Initial risk groups' lowest cut-point at the expense of a small decrease in sensitivity. The sensitivity, specificity, PPV and NPV of the other two corresponding cut-points were similar. Due to the comparable performance of the Initial and Simplified risk groups in distinguishing between individuals with normal and those with abnormal glucose levels, as well as in matching the LSA risk groups and screening decisions the Simplified risk groups are advocated. The Simplified risk groups have the benefit over the Initial risk groups of being easier to remember for both those receiving their risk group and healthcare professionals explaining the meaning of each risk group to patients.

The lowest cut-point of the Simplified risk groups, ≥0.105, had very high sensitivity showing only a small proportion people with abnormal glucose would be classified as low risk under the Simplified risk groups. The highest cut-point of Simplified risk groups, ≥0.305, had high specificity meaning only a small proportion of individuals with normal glucose would be classified as very high

risk. The cut-point of ≥0.155 which decides the screening decision had high sensitivity indicating most individuals with abnormal glucose would be offered a blood test. The modest specificities, 49.4% in ADDITION-Leicester and 47.5% in STAR show that around half of individuals with normal HbA1c measurements would be invited to have a blood test. In this context, having a high sensitivity and modest specificity is acceptable as reducing false negatives is more important than reducing false positives (108). The PPVs and NPVs show that slightly more than three of ten individuals offered blood test would have abnormal HbA1c measurements and a little less than nine in ten not offered would have normal HbA1c measurements.

In addition to performing well in distinguishing between individuals with normal and those with abnormal glucose levels, it is desirable that the chosen LPRS categories have high agreement with the existing LSA categories. This is because the LSA risk score is widely used (194), thus if the LPRS categories are implemented nationally a proportion of individuals will receive two communications about their risk from the two different risk scores. In order for individuals to receive a consistent message from the two RATs, the risk groups for the two have to match, where this is not the case it is best if they are similar to one another. Both the Initial and Simplified risk groups had moderate levels of agreement with the LSA risk groups, with agreement in the external dataset being 65.6% for the Initial risk groups and 60.8% for the Simplified risk groups. Reassuringly, in the cases of both the Initial and Simplified risk groups only a handful of individuals were classified as a risk group which differed by two categories from their LSA risk groups. This indicates that the vast majority of individuals not given the same risk groups as for LSA had a risk group only one category different using either the Initial or Simplified LPRS risk groups, and thus would receive a similar message about their risk from the two scores. Furthermore, both the Initial and Simplified LPRS screening decisions had substantial agreement with LSA screening decisions, with 82.7% and 85.1% agreement respectively. The screening decision matching that of the LSA's in most cases is of importance. In addition to determining whether a blood test is offered, the screening decision is likely to be the most influential part of risk communication in terms of how an individual understands their risk.

The STAR dataset had a higher prevalence of the outcome, 29.0%, than the ADDITION-Leicester dataset, 22.6%. The higher prevalence may be due to the inclusion criterion of the STAR study of individuals having a risk factor for diabetes. Due to this inclusion criterion the proportion of individuals classified as low or moderate risk by the Initial LPRS and the Simplified LPRS risk groups was slightly lower in the external data than the internal data; while there was a marked drop in these proportions for the LSA risk groups.

7.5.2 Strengths and weaknesses of analysis

One strength of this analysis was the use of two multi-ethnic datasets comprising of individuals living in both urban and rural locations. This meant one dataset could be used to develop the two sets of risk groups, while the second dataset was used to assess their external performance. A further strength is that the datasets contained the whole age in which screening for abnormal glucose is advocated, since they were both collected with screening for abnormal glucose in mind. Finally, the sensitivity analysis carried out ensured the findings were consistent whether HbA1c or FPG was used to define abnormal glucose. This means the results are relevant to general practices whether they currently use HbA1c or FPG as the second stage of screening advised by NICE.

The datasets had the weakness that they were only comprised of individuals from Leicestershire, to help overcome this limitation the Simplified screening decision cut-point of ≥0.155 will be assessed in a nationally representative dataset in the next chapter of this thesis. Another weakness is STAR required individuals to have at least one risk factor for diabetes. This means that the proportion of individuals in different risk groups will be slightly skewed compared to the general population, with a little less being in the low and moderate risk groups and a few more in the high and very high risk groups. The proportions observed in the different risk groups for the ADDITION-Leicester dataset are likely to give a better estimation of the proportions yielded nationally if implemented in practice. Although the proportion of individuals classified as high and very high may be slightly increased due to a higher proportion of individuals identifying as being BME in this dataset than nationally. For this reason, the proportion screened as positive using the cut-point of ≥0.155 will be

calculated in the next chapter to give a reliable indication of the proportion that would require to be offered a blood test if implemented nationally.

7.6 Conclusion and implications for this thesis

The results from this chapter support the use of the Simplified LPRS risk groups across general practices. The use of the Simplified risk groups would lead to approximately 56.5% of 40- 75 year olds being classified as high or very high risk nationally. Screening all these individuals at once may be unfeasible for general practices, particularly those with a disproportionately large number of individuals at high or very high risk. An achievable screening strategy for practices may be to invite individuals identified as very high risk, roughly 17% nationally, to targeted blood test appointments; and opportunistically offer blood tests to individuals identified as high risk when they visit the practice for another consultation, as this will require less resources to implement. Although it should be noted to identify around 75% of the individuals with abnormal glucose it is necessary to screen all individuals deemed to be high or very high risk by the Simplified LPRS risk groups.

Chapter 8: External national validation of the Leicester Self-Assessment and Leicester Practice Risk Scores using data from the English Longitudinal Study of Ageing

8.1 Chapter Outline

This chapter assesses the validity of the Leicester Self-Assessment (LSA) and Leicester Practice Risk Score (LPRS) to identify individuals who develop diabetes prospectively using a nationally representative dataset. Evaluating the risk assessment tools (RATs) when used alone, as well as the performance of using each of the RATs as the first stage in a two-staged screening approach, with the second stage being a blood test for those categorised as high or very high risk by the RAT being used. Additionally, the validity of the RATs for the outcome of prevalent non-diabetic hyperglycaemia (NDH) or undiagnosed Type 2 diabetes mellitus (T2DM) in the nationally representative dataset is assessed.

The work in this chapter has been:

- Orally presented:
  Barber SR, Dhalwani NN, Davies MJ, Khunti K, Gray LJ.. 'Prospective validation of The Leicester/Diabetes UK Risk Assessment for diagnosis of Type 2 diabetes.' Diabetes UK Professional Conference 2016. Glasgow, UK. 2nd-4th March 2016. (A71, P254)

8.2 Introduction

The LSA and LPRS are non-invasive RATs which have been detailed in section 1.4.2. To summarise, the RATs were developed using a cross-sectional multi-ethnic Leicestershire dataset to identify individuals who currently have abnormal glucose levels, either NDH or undiagnosed diabetes (46). The LSA has been developed so that it can be calculated by the individual themselves or by a lay person with the individual present; while the LPRS was designed to be calculated by a computer using routine information stored in general practices' databases. A cut-point of ≥16 for the LSA score is recommended, and used in clinical practice, for identifying individuals who are at high risk and require a blood test (51). While the work in Chapter 7 led to the recommendation that individuals with LPRS ≥0.155 be offered a blood test.

To date both RATs have only been validated for cross-sectional outcomes, predominately using Leicester based data (25,45,51). This chapter uses data from the English Longitudinal Study of Aging (ELSA) (91), a nationally representative dataset of individuals 50 years and over to externally validate the RATs. Firstly, this chapter evaluates the cross-sectional performance of the RATs in identifying individuals who currently have abnormal glucose levels, either those with undiagnosed diabetes or those with NDH or undiagnosed diabetes. The discrimination and calibration of the two RATs for these outcomes are calculated; as well as predictive diagnostics of the cut-points recommended for deciding whether individuals should be offered a blood test by each RAT.

The concept of identifying individuals who currently have an abnormal glucose level is that individuals will be identified earlier along the diabetes pathway and thus be given interventions to prevent or delay their progression to diabetes. However, no validation for the outcome of developing diabetes in the years following the RATs' calculation using longitudinal data has been carried out to date for either of the RATs. Therefore, the discrimination and calibration of the two RATs in detecting the incidence of T2DM within four and eight years are calculated. Predictive diagnostics of the cut-points recommended for deciding whether individuals should be offered a blood test by each RAT are also calculated for these binary longitudinal outcomes.

The National Institute for Health and Care Excellence (NICE) recommend identifying individuals between the age of 40 and 75 years with NDH or undiagnosed T2DM using a two-stage screening approach involving a non-invasive RAT followed by a blood test for individuals with a high risk score (28). Figure 8.1 shows the screening programme recommended in the NICE Public Health Guidelines 38 (28). Individuals with a high risk score followed by an abnormal glucose measurement are either entered into the diabetes pathway, if further investigation confirms diabetes, or are invited to take part in an intensive lifestyle-change programme. Thus individuals with a high risk score followed by an abnormal glucose measurement are defined as screening positive by the two-stage screening programme. This chapter evaluates the performance of the two staged screening programme with the LSA and LPRS as the first stage of screening in identifying the individuals who prospectively develop T2DM. Thus assessing the utility of implementing these RATs in practice, as part of the two-stage screening approach in identifying the individuals who without intervention would go onto develop diabetes in the years that follow.

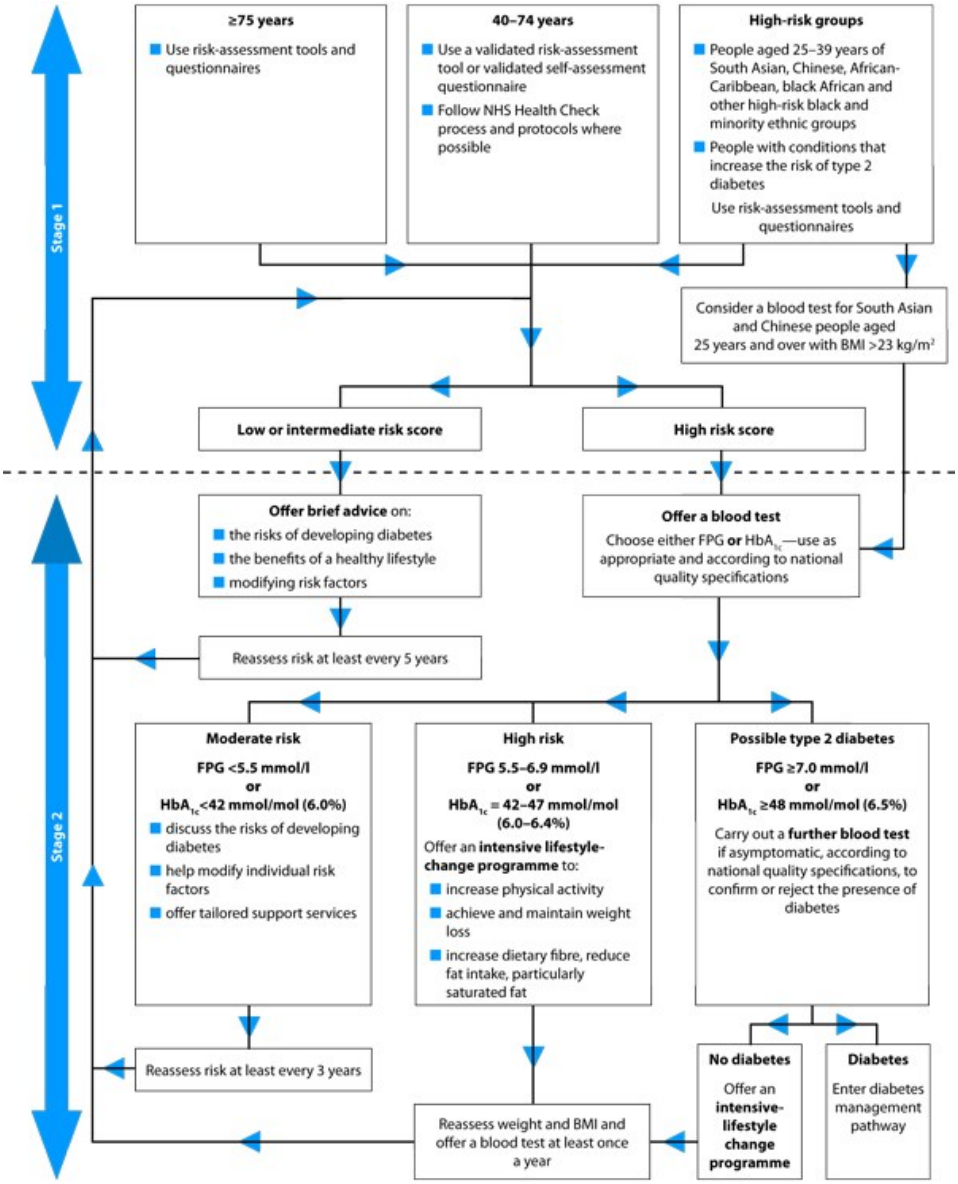


Figure 8.1 NICE guidelines on identifying and managing risk of T2DM (28)

8.3 Methods

8.3.1 Dataset

Data were taken from ELSA, a nationally representative dataset of people aged 50 years and older which has been detailed in Chapter 2. ELSA collects data from participants every two years, each round of data collection is known as a wave. Participants are asked to complete a questionnaire every wave, with nurse visits being conducted every other wave (every four years) to collect further information, such as blood test measurements. The work in this chapter uses Wave 2 (conducted in 2004 and 2005) as the baseline, as this is the first wave which included nurse visits and therefore blood test measurements. The most recent data available from ELSA at the time of carrying out this work is from Wave 6 (conducted in 2012 and 2013) meaning a follow-up period of eight years was available for the longitudinal analyses in this chapter.

Due to high levels of missing data the main analyses presented in this chapter were carried out on multiply imputed data. Sensitivity analyses using complete-case data were also carried out, results of which are reported in Appendix C. The multiple imputation used fully conditional specification (FCS) to give each missing value 50 imputations. The FCS specified each variable with missing values a conditional distribution using the risk factors of both RATs and the following variables: glycated haemoglobin A1c (HbA1c) at baseline, HbA1c at four year follow-up, HbA1c at eight year follow-up, fasting plasma glucose (FPG) at baseline, FPG at four year follow-up, FPG at eight years, weight, height, systolic blood pressure, diastolic blood pressure, current smoker (yes/no) and cholesterol. Continuous variables which did not follow a normal distribution were log transformed before imputation was carried out and then back transformed along with the imputed values afterwards. Continuous variables which still did not follow a normal distribution after log-transformation were imputed using bootstrapping to overcome this issue.

8.3.2 Score calculation

The two RATs were calculated for individuals aged between 50–75 years old (in Wave 2), who did not have diagnosed diabetes by Wave 2. Table 8.1 shows the ELSA variables used to calculate the LSA and LPRS. Due to the absence of any information on family history of diabetes in Wave 2, the family history variable was imputed from Wave 6 (2012/13) using whether an individual has/had a parent who has/had diabetes at this point instead of whether an individual has/had a first degree family member who has/had diabetes at baseline. All other risk factors of the two RATs were all variables recorded in Wave 2.

Table 8.1 ELSA variables used to calculate LSA and LPRS

Risk factor	ELSA variable used for LSA	ELSA variable used for LPRS
Age	Age at Wave 2	Age at Wave 2
Sex	Sex at Wave 2	Sex at Wave 2
Ethnicity	Ethnicity at Wave 2	Ethnicity at Wave 2
Family history of T2DM	Parents' history of diabetes at Wave 6	Parents' history of diabetes at Wave 6
Waist circumference (cm)	Waist circumference at Wave 2	N/A
Body mass index (kg/m^2)	BMI at Wave 2	BMI at Wave 2
High blood pressure	Reported been diagnosed with high blood pressure before Wave 2	Reported antihypertensive medication use at Wave 2

8.3.3 Analyses

Stata 13 (146) was used to carry out the analyses. For reasons which were outlined in the introduction, the performance of the following has been analysed:

- Each of the RATs for binary diabetes-related outcomes at baseline.
- Each of the RATs for the development of diabetes within the follow-up period for those free from diabetes at baseline.
- The two-staged screening programme (LSA or LPRS followed by a blood test at baseline for those above the cut-point of the RAT being used) for the development of diabetes within the follow-up period for those free from diabetes at baseline.

The definitions of the various baseline outcomes, for which the performance of the two RATs to detect has been assessed, are shown in Table 8.2. FPG and HbA1c were used to define the binary outcomes of both undiagnosed diabetes alone and NDH or undiagnosed diabetes. Two different cut-points are used to define abnormal glucose by FPG, firstly ≥5.5mmol/l which is the cut-point used in NICE PH38 to defined high risk by FPG; and secondly ≥6.1mmol/l which is the cut-point above which an individual has either impaired fasting glucose (IFG) or T2DM.

Table 8.2 Definition of the various baseline outcomes which the risk assessment tools were assessed for detecting

	Definition of outcome
Undiagnosed diabetes	FPG ≥ 7.0mmol/l
	HbA1c ≥ 6.5%
NDH or undiagnosed diabetes	FPG ≥ 5.5mmol/l
	FPG ≥ 6.1mmol/l
	HbA1c ≥ 6.0%

The longitudinal outcomes which were used to assess the ability of the RATs, alone and followed by a baseline blood test, to detect those individuals who go on to develop diabetes in the years following their risk assessment were:

- Self-reported doctor diagnosed diabetes within eight years
- FPG ≥ 7.0mmol/l at four year follow-up (Wave 4: 2008/2009) or self-reported doctor diagnosed diabetes within four years
- HbA1c ≥ 6.5% at four year follow-up (Wave 4:2008/2009) or self-reported doctor diagnosed diabetes within four years
- FPG ≥ 7.0mmol/l at eight year follow-up (Wave 6: 2012/2013) or self-reported doctor diagnosed diabetes within eight years
- HbA1c ≥ 6.5% at eight year follow-up (Wave 6:2012/2013) or self-reported doctor diagnosed diabetes within eight years

For the longitudinal analyses doctor diagnosed diabetes was considered the primary outcome, since the other outcomes do not confirm diabetes on their own; instead requiring a confirmatory blood test.

For each of the outcomes evaluated in this chapter, the area under the receiver operator curve (AUROC) of the RATs or the two-stage screening process was calculated, along with its 95% confidence interval (CI), to assess the discrimination of the score in each case. Each RAT has an associated probability of having the outcome of interest. The calibration of the RATs was assessed for the various outcomes by calculating the Brier score using these associated probabilities. The outcome index variance, the Brier score yielded by predicting the prevalence of the outcome for each individual is displayed to allow a comparison to the non-informative model to be made. For the two-stage screening process, calculating the Brier score was not possible since there are not any associated probabilities for the different groupings.

The prevalence of the both the cross-sectional and longitudinal outcomes were calculated across the risk groupings of each RAT. The following predictive diagnostics were calculated for the RATs' recommended cut-points (LSA ≥16, LPRS≥0.155) for both the cross-sectional and longitudinal outcomes:

- Specificity
- Sensitivity
- Positive Predictive Value (PPV)
- Negative Predictive Value (NPV)
- Percentage correctly classified
- Percentage classified as high risk

These predictive diagnostics for the longitudinal outcomes were also calculated for the two-stage screening programme, with individuals being defined as having screened positive if their RAT was greater than or equal to the cut point (LSA ≥16/ LPRS≥0.155) and their resulting blood test was abnormal (FPG ≥5.5mmol/l, HbA1c ≥6.0%). Separate analyses were carried out to assess the impact of the blood test used with both FPG and HbA1c used to define blood glucose status at baseline. For ease of reading in this Chapter, the term blood glucose will be used to refer to both FPG and HbA1c measurements. Finally, the proportion of individuals diagnosed with diabetes within eight years was calculated for the following three groups: individuals with a low or moderate risk score; individuals with a high risk score and normal baseline glucose measurement; individuals with a high risk score and abnormal baseline glucose measurement.

8.4 Results

Of the 9,432 individuals that took part in ELSA Wave 2 there were 6,778 individuals in the population of interest for the analyses carried out. The main analyses reported in this chapter included all 6,778 individuals in the population of interest since multiple imputation was used for missing values of risk factors and outcomes.

Table 8.3 gives the summary statistics of the observed data compared to the full dataset used in the main analyses with missing values multiply imputed. The proportion with each risk factor and outcome is also given. Three thousand nine hundred and two (57.6%) had complete data for all the risk factors in the LSA; while 3,935 (58.1%) had complete data for all the risk factors in the LPRS. The number with complete data for each analysis is given in Appendix C where results of the sensitivity analyses are displayed.

Table 8.3 Summary statistics of the risk factors and outcomes observed compared to data multiple imputed in population of interest (n=6,778)

	Observed data	Proportion with missing	Data with missing values multiply imputed
Age (years)	62.3 (6.84)	0%	62.3 (6.84)
50-59	41.0%	-	41.0%
60-69	39.7%	-	39.7%
70-75	19.3%	-	19.3%
BMI (Kg/m^2)	27.9 (4.79)	18.6%	27.9 (4.76)
<25	27.4%	-	23.2%
≥25 & <30	44.0%	-	44.8%
≥30 & <35	20.9%	-	23.4%
≥35	7.6%	-	8.7%
Ethnicity (White European)	97.6%	0%	97.6%
Family History of diabetes*	14.6%	31.9%	14.4%
High blood pressure or antihypertensive medication use	38.5%	0%	38.5%
Antihypertensive medication use	12.5%	0%	12.5%
Waist circumference (cm)	95.2(13.1)	17.8%	95.1 (12.7)
<90	28.3%	-	34.2%
90-99	24.2%	-	30.4%
100-109	18.2%	-	22.6%
≥110	11.3%	-	12.8%
Sex (Male)	44.5%	0%	44.5%
Baseline FPG (mmol/l)	4.96 (0.776)	53.3%	4.97 (0.740)
Baseline HbA1c (%)	5.47 (0.490)	34.4%	5.47 (0.480)
FPG at four year follow-up (mmol/l)	4.87 (0.731)	65.5%	4.93 (0.731)
HbA1c at four year follow-up (%)	5.80 (0.515)	50.9%	5.82 (0.519)
FPG at eight year follow-up (mmol/l)	5.38 (0.833)	74.6%	5.50 (0.903)
HbA1c at eight year follow-up (%)	6.08 (0.613)	56.3%	5.89 (0.631)
Diagnosed diabetes at eight year follow-up	8.41%	31.3%	8.03%

Values shown as mean (SD), unless stated
*(Mother or Father who had diabetes by **Wave 6**)

8.4.1 LSA risk score for cross-sectional outcomes

Table 8.4 shows the LSA had higher AUROCs for detecting baseline diabetes range outcomes than baseline NDH or diabetes range outcomes. The Brier scores were markedly higher than their associated outcome index variances for the diabetes range outcomes. For both the diabetes range outcomes and NDH or diabetes range outcomes the AUROCs were highest for the HbA1c defined outcome.

Table 8.4 Discrimination and calibration of LSA risk score for various binary cross-sectional outcomes in ELSA dataset

Outcome	Prevalence	AUROC (95% confidence Interval)	Brier score (outcome index variance)
Diabetes: FPG ≥ 7.0mmol/l	1.32% (0.877, 1.76)	0.710 (0.646, 0.774)	0.0657 (0.0130)
Diabetes: HbA1c ≥ 6.5%	2.30% (1.86, 2.75)	0.732 (0.686, 0.780)	0.0689 (0.0225)
NDH or diabetes: FPG ≥ 5.5mmol/l	17.7% (16.2, 19.1)	0.627 (0.604, 0.649)	0.146 (0.146)
NDH or diabetes: FPG ≥ 6.1mmol/l	5.91% (4.90, 6.92)	0.669 (0.635, 0.704)	0.0860 (0.0556)
NDH or diabetes: HbA1c ≥ 6.0%	9.57% (8.73, 10.4)	0.680 (0.654, 0.707)	0.101 (0.0865)

63.3% (95% CI: 62.1, 64.5) of the individuals included in the analysis were classified as high risk using the LSA with a cut-point of ≥16. Table 8.5 displays using LSA ≥16 for detecting individuals with a blood test measure in the diabetes range at baseline had a high sensitivity but low specificity. The NPVs were very high, however the PPVs were low. Using the same cut-point for detecting individuals with a blood test measure in the NDH or diabetes range at baseline led to a drop in sensitivity, marginally increased specificity, higher PPV and slightly lower NPV for various definitions of the outcome.

Table 8.5 Predictive diagnostics of LSA, with cut-point ≥16, for various binary cross-sectional outcomes in ELSA dataset

Outcome	Sensitivity	Specificity	PPV	NPV	Correctly Classified
Diabetes: FPG ≥ 7.0mmol/l	91.5 (82.7, 100)	37.1 (35.9, 38.3)	1.9 (1.2, 2.6)	99.7 (99.4, 100.0)	37.8 (36.6, 39.1)
Diabetes: HbA1c ≥ 6.5%	89.7 (83.9, 95.5)	37.4 (36.1, 38.6)	3.3 (2.6, 3.9)	99.4 (99.0, 99.7)	38.6 (37.3, 39.8)
NDH or diabetes: FPG ≥ 5.5mmol/l	77.0 (74.1, 79.9)	39.7 (38.3, 41.1)	21.5 (19.7, 23.4)	88.9 (87.3, 90.6)	46.3 (44.8, 47.7)
NDH or diabetes: FPG ≥ 6.1mmol/l	83.2 (78.5, 87.9)	38.0 (36.7, 39.3)	7.8 (6.4, 9.1)	97.3 (96.4, 98.2)	40.7 (39.3, 42.0)
NDH or diabetes: HbA1c ≥ 6.0%	83.3 (79.8, 86.8)	38.9 (37.6, 40.2)	12.6 (11.4, 13.8)	95.7 (94.7, 96.6)	43.1 (41.8, 44.4)

Values given as % (95% CI)

Figure 8.2 and Figure 8.3 show the prevalence of all the baseline outcomes increased in each LSA risk group.

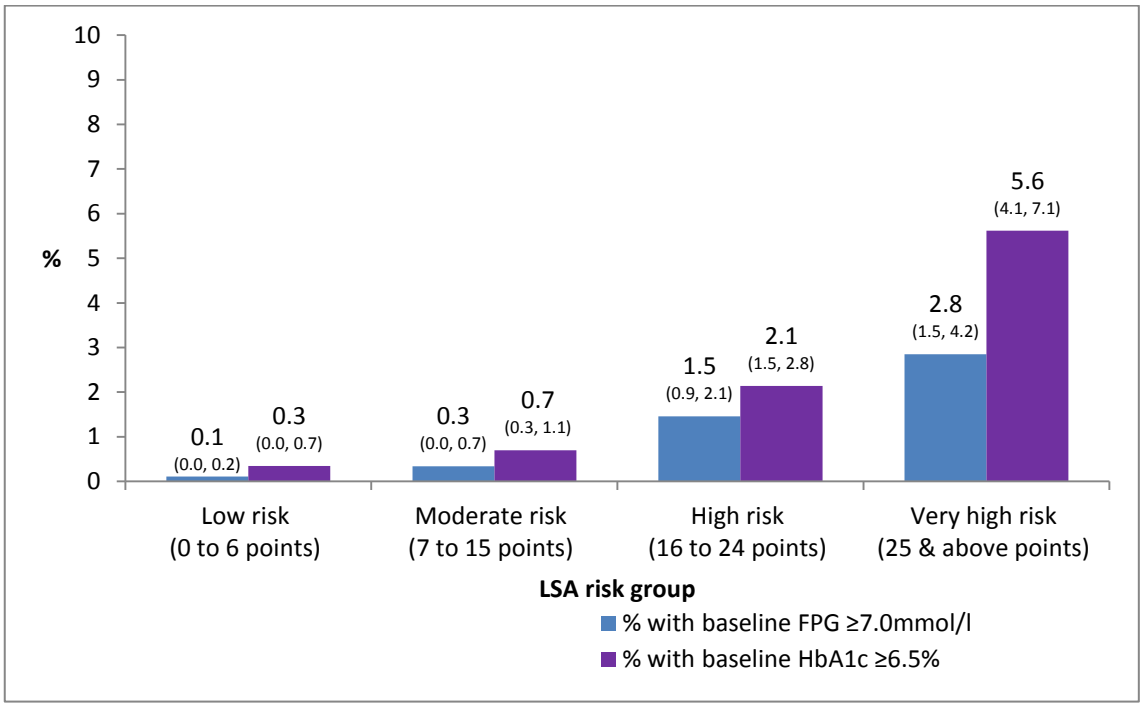


Figure 8.2 Percentage of individuals in each LSA risk group with FPG ≥7.0mmol/l and HbA1c ≥6.5% at baseline

Data displayed as % (95% CI)

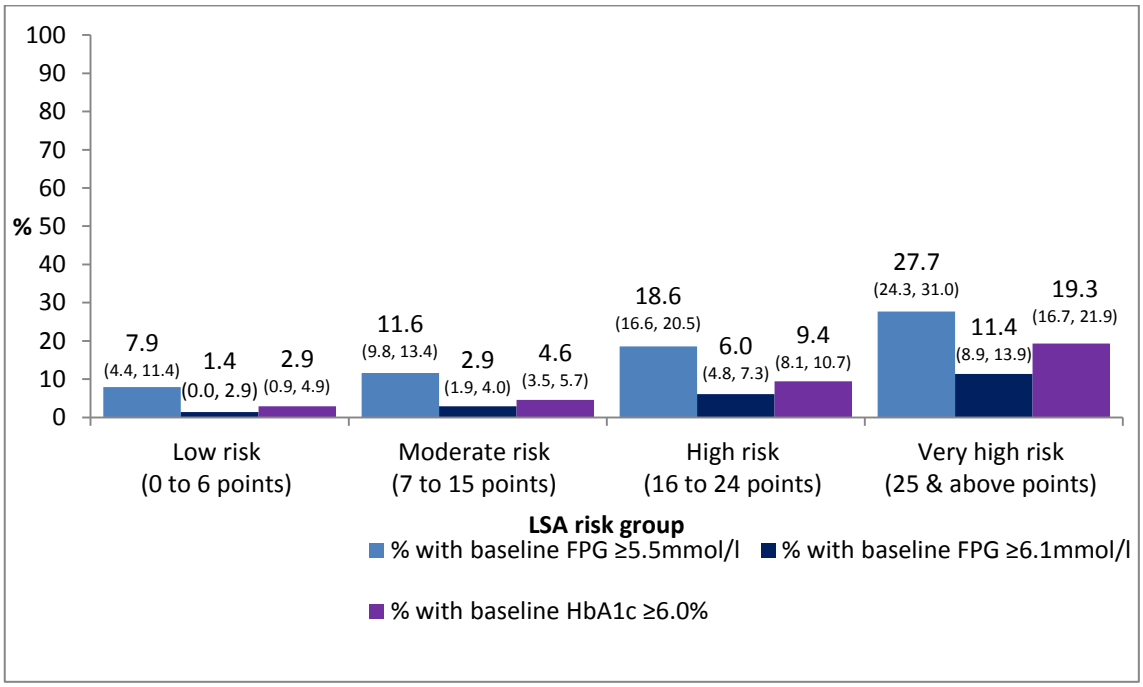


Figure 8.3 Percentage of individuals in each LSA risk group with FPG ≥5.5mmol/l, FPG ≥6.1mmol/l and HbA1c ≥6.0% at baseline

Data displayed as % (95% CI)

8.4.2 LSA risk score for longitudinal outcomes

Table 8.6 shows that 8.03% of individuals reported being diagnosed with diabetes by a doctor within eight years of baseline. The prevalence of individuals that had a measurement in the diabetes range was between 5.56% and 13.3% for the two different blood tests at four and eight years after baseline. The AUROC ranged from 0.699 to 0.726, with the AUROCs for the HbA1c outcomes being marginally lower than the AUROCs for the FPG outcomes. All Brier scores were higher than their associated outcome index variances.

Table 8.6 Discrimination and calibration of LSA for various binary longitudinal diabetes outcomes in ELSA dataset

Outcome	**Prevalence**	**AUROC** (95% confidence interval)	**Brier score** (Outcome index variance)
Diabetes: FPG≥ 7.0mmol/l at Wave 4 or doctor diagnosed diabetes within four years	5.56% (4.86, 6.26)	0.726 (0.696, 0.755)	0.0812 (0.0525)
Diabetes: HbA1c ≥ 6.5% at Wave 4 or doctor diagnosed diabetes within four years	9.41% (8.52, 10.3)	0.718 (0.693, 0.742)	0.0979 (0.0852)
Diabetes: FPG≥ 7.0mmol/l at Wave 6 or doctor diagnosed diabetes within eight years	10.3% (9.02, 11.6)	0.702 (0.678, 0.726)	0.103 (0.0923)
Diabetes: HbA1c ≥ 6.5% at Wave 6 or doctor diagnosed diabetes within eight years	13.3% (12.1, 14.6)	0.699 (0.674, 0.723)	0.117 (0.115)
Doctor Diagnosed within eight years	8.03% (7.25, 8.82)	0.717 (0.691, 0.742)	0.0926 (0.0739)

Table 8.7 shows using the LSA, with cut-point ≥16, produced high sensitivity and very high NPV, though it yielded low specificity and low PPV for the various binary longitudinal outcomes. The extremely high NPVs indicate the LSA with this cut-point is good for ruling out individuals who are unlikely to develop T2DM in the following eight years.

Table 8.7 Predictive diagnostic of LSA, with cut-point ≥16, for various binary longitudinal outcomes in ELSA dataset

Outcome	Sensitivity	Specificity	PPV	NPV	Correctly Classified (%)
Diabetes: FPG≥ 7.0mmol/l at Wave 4 or doctor diagnosed diabetes within four years	88.3 (84.1, 92.4)	38.7 (37.4, 40.0)	7.8 (6.7, 8.9)	98.3 (97.6, 98.9)	40.3 (39.0, 41.7)
Diabetes: HbA1c ≥ 6.5% at Wave 4 or doctor diagnosed diabetes within four years	87.3 (83.8, 90.8)	39.4 (38.1, 40.7)	13.0 (11.8, 14.3)	96.8 (95.8, 97.7)	43.5 (42.2, 44.8)
Diabetes: FPG≥ 7.0mmol/l at Wave 6 or doctor diagnosed diabetes within eight years	85.4 (82.1, 88.6)	39.3 (37.9, 40.6)	13.9 (12.2, 15.6)	95.9 (94.8, 97.0)	44.0 (42.6, 45.5)
Diabetes: HbA1c ≥ 6.5% at Wave 6 or doctor diagnosed diabetes within eight years	84.9 (81.9, 87.9)	39.4 (38.1, 40.7)	17.9 (16.2, 19.5)	94.5 (93.3, 95.8)	46.0 (44.6, 47.4)
Doctor Diagnosed within eight years	87.9 (84.6, 91.1)	38.9 (37.6, 40.2)	11.2 (10.0, 12.3)	97.3 (96.6, 98.1)	42.8 (41.5, 44.1)

Values given as % (95% CI)

Figure 8.4 and Figure 8.5 show proportion of individuals with each longitudinal outcomes increases in each LSA risk group.

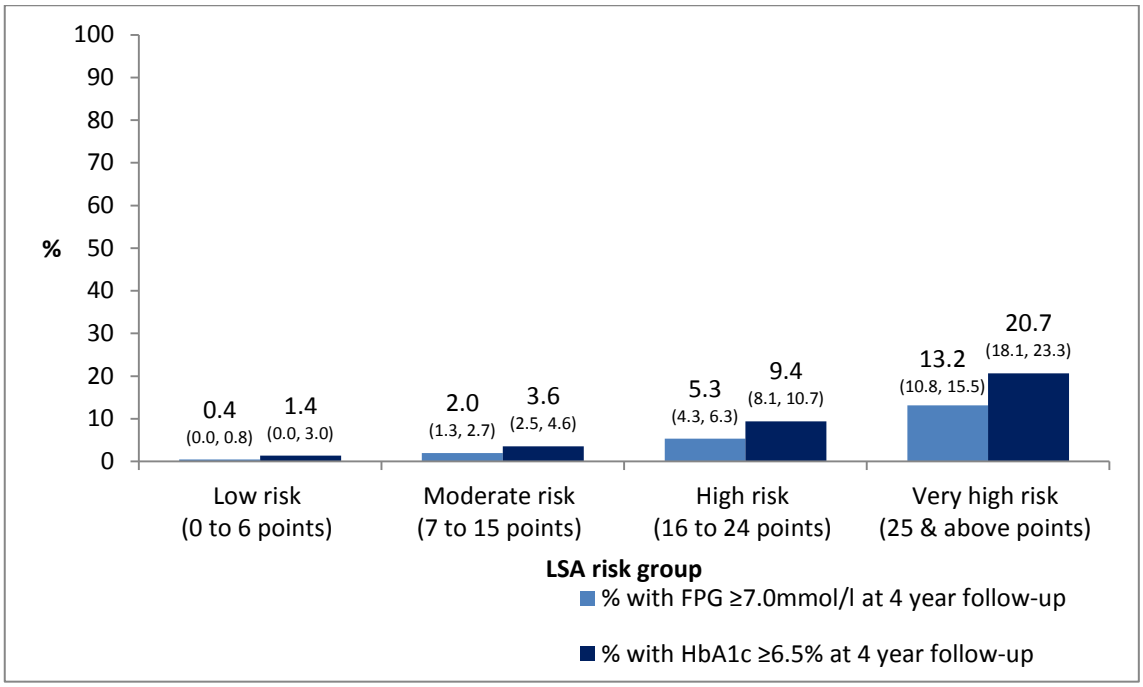


Figure 8.4 Percentage of individuals in each LSA risk group with FPG ≥7.0mmol/l and HbA1c ≥6.5% at four year follow-up

Data displayed as % (95% CI)

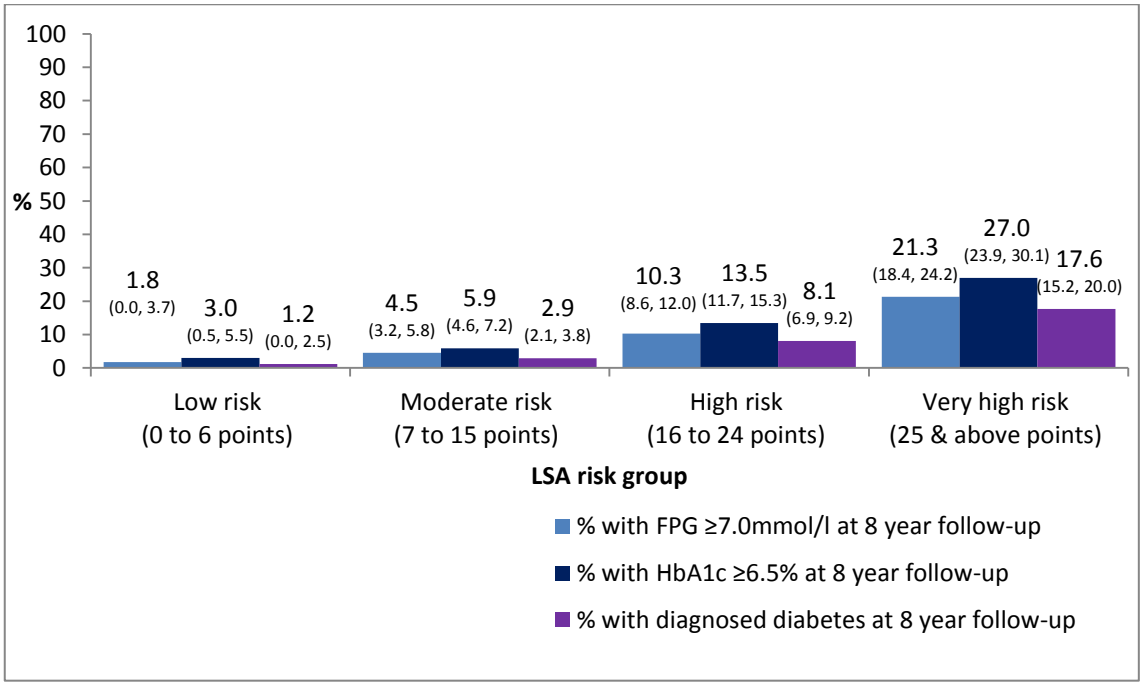


Figure 8.5 Percentage of individuals in each LSA risk group with FPG ≥7.0mmol/l, HbA1c ≥6.5% and diagnosed diabetes at eight year follow-up

Data displayed as % (95% CI)

8.4.3 Two-staged approach with LSA as first stage for longitudinal outcomes

Table 8.8 shows the two-stage screening programme with the LSA as the first stage gave AUROCs ranging from 0.721 to 0.761. Using HbA1c in the second stage resulted in the best discrimination for outcome.

Table 8.8 Discrimination of two-stage screening programme, with LSA as first stage and various blood tests as second stage, for binary longitudinal outcome of doctor diagnosed diabetes within eight years in ELSA dataset

Blood screening test used	AUROC
FPG (with NDH defined as FPG≥5.5mmol/l)	0.755 (0.725, 0.786)
FPG (with NDH defined as FPG≥6.1mmol/l)	0.721 (0.693, 0.750)
HbA1c	0.761 (0.734, 0.787)

Table 8.9 shows the specificity and NPV of the two-stage screening programme's decision were both high for different blood tests and cut-points analysed for stage two. Between 4.9% and 13.6% of individuals were classified as high risk depending on the blood test and cut-point used. Using HbA1c ≥6.0% as the cut-point in the second stage yielded the best combination of predictive diagnostics with sensitivity and PPV both being a little under 50%.

Table 8.9 Predictive diagnostics of screening decision of the two-stage screening programme, with LSA as first stage and various blood tests as second stage, for binary longitudinal outcome of doctor diagnosed diabetes within 8-years in ELSA dataset

Blood test used	Sensitivity	Specificity	PPV	NPV	Correctly Classified	Classified as high risk
FPG (with NDH defined as FPG≥5.5mmol/l)	52.5 (45.8, 59.2)	89.8 (88.7, 90.9)	31.0 (26.7, 35.4)	95.6 (94.8, 96.3)	86.8 (85.6, 88.0)	13.6 (12.4, 14.8)
FPG (with NDH defined as FPG≥6.1mmol/l)	32.0 (25.5, 38.4)	97.4 (96.8, 98.1)	52.3 (43.7, 60.9)	94.3 (93.5, 95.0)	92.2 (91.3, 93.1)	4.9 (4.0, 5.8)
HbA1c	47.0 (41.7, 52.2)	95.4 (94.8, 96.1)	47.3 (41.9, 52.8)	95.4 (94.7, 96.0)	91.5 (90.7, 92.4)	8.0 (7.2, 8.8)

Values given as % (95% CI)

Figure 8.6 displays the prevalence of doctor diagnosed diabetes within eight years for the different risk groupings of two-staged baseline screening with LSA as the first stage. It shows that the groupings which are below the cut-off for receiving an intensive lifestyle intervention (LSA<16 and LSA≥16 & blood glucose in the normal range) had a low prevalence of the outcome; lower than seen for the whole dataset. These two groups contain between 86.5% and 95.1% of individuals depending on the blood test and cut-point used for the second stage of screening.

The group of individuals with LSA ≥16 and NDH has been combined with the group with LSA ≥16 and a glucose measurement in the diabetes range, meaning the prevalence is shown for individuals with a positive screening decision from the full screening programme. Almost half the individuals screening positive using either HbA1c ≥6.0% or FPG ≥6.1mmol/l as the cut-point in the second stage were diagnosed with diabetes within eight years; this dropped to around a third when FPG ≥5.5mmol/l was used as the cut-point in the second stage.

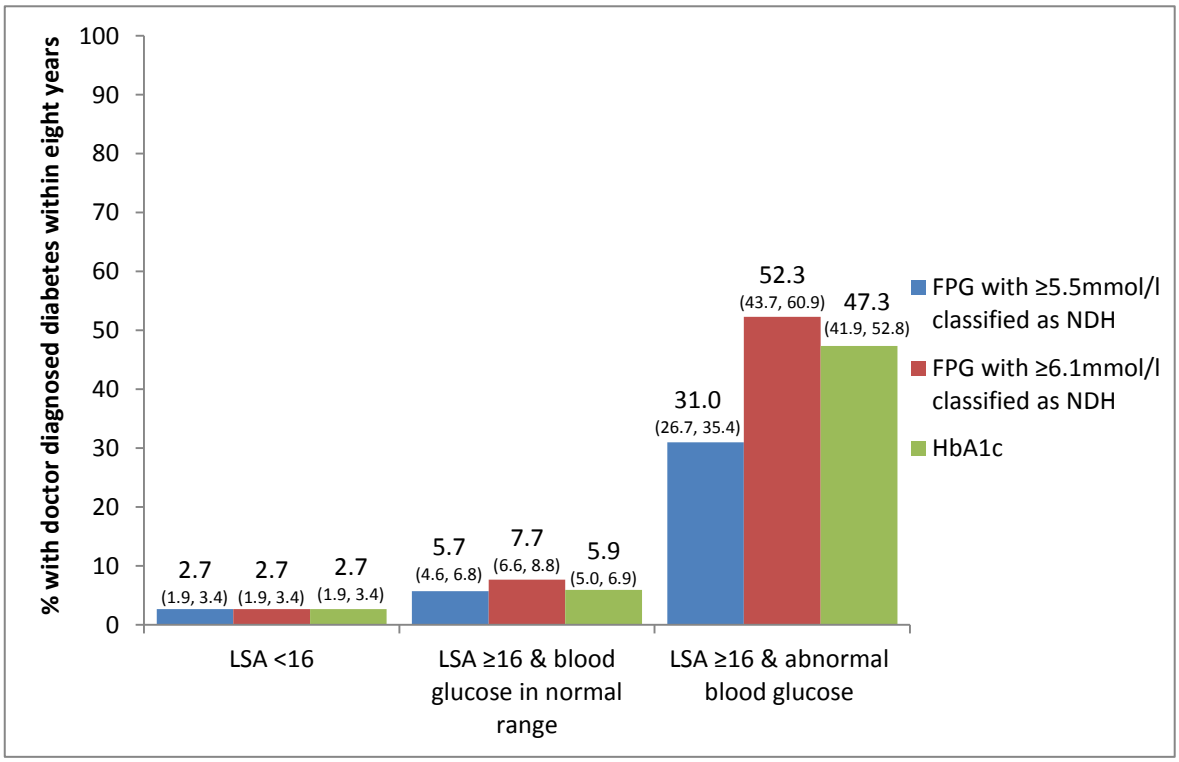


Figure 8.6 Percentage of individuals from different groupings of two-stage (with LSA risk score used at stage one) baseline screening being diagnosed with diabetes by a doctor within eight years for various baseline blood tests

Data displayed as % (95% CI)

8.4.4 LPRS for cross-sectional outcomes

Table 8.10 shows the AUROC of the LPRS for detecting individuals with blood glucose in the diabetes range at baseline was higher than the AUROC of the LPRS for detecting those with blood glucose in the abnormal range. The Brier scores were noticeably higher than their associated outcome index variances for the diabetes range outcomes. For both the diabetes range outcomes and NDH or diabetes range outcomes the AUROC was highest for the HbA1c defined outcome.

Table 8.10 Discrimination and calibration of LPRS score for various binary cross-sectional outcomes in ELSA dataset

Outcome	**Prevalence**	**AUROC** (95% confidence interval)	**Brier score** (Outcome index variance)
Diabetes: FPG ≥7.0mmol/l	1.32% (0.877, 1.76)	0.678 (0.613, 0.743)	0.0428 (0.0130)
Diabetes: HbA1c ≥6.5%	2.30% (1.86, 2.75)	0.723 (0.674, 0.771)	0.0474 (0.0225)
NDH or diabetes: FPG ≥5.5mmol/l	17.7% (16.2, 19.1)	0.613 (0.590, 0.637)	0.143 (0.146)
NDH or diabetes: FPG ≥6.1mmol/l	5.91% (4.90, 6.92)	0.650 (0.615, 0.684)	0.0697 (0.0556)
NDH or diabetes: HbA1c ≥6.0%	9.57% (8.73, 10.4)	0.665 (0.639, 0.0.691)	0.0901 (0.0865)

46.0% (95% CI: 44.8, 47.3) of the individuals included in the analysis were classified as high risk using the LPRS with a cut-point of ≥0.155. Using LPRS≥0.155 for detecting individuals with a blood glucose measure in the diabetes range at baseline had fairly high sensitivity and reasonable specificity for the various blood tests. The NPVs were very high, however the PPVs were low. Using the same cut-point for detecting individuals with a blood glucose measure in the NDH or diabetes range at baseline led to a drop in sensitivity, very marginally greater specificity, higher PPV and slightly lower NPV for various blood tests and cut-points.

Table 8.11 Sensitivity, specificity, PPV, NPV, proportion correctly classified and proportion classified as high risk of LPRS, with cut-point ≥0.155, for various binary cross-sectional outcomes in ELSA dataset

Outcome	Sensitivity	Specificity	PPV	NPV	Correctly Classified
Diabetes: FPG ≥7.0mmol/l	70.0 (56.9, 83.1)	54.3 (53.0, 55.6)	2.0 (1.2, 2.8)	99.3 (98.9, 99.6)	54.5 (53.2, 55.8)
Diabetes: HbA1c ≥6.5%	77.4 (69.0, 85.7)	54.7 (53.4, 56.0)	3.9 (3.0, 4.7)	99.0 (98.6, 99.4)	55.2 (53.9, 56.5)
NDH or diabetes: FPG ≥5.5mmol/l	59.6 (56.1, 63.0)	56.9 (55.4, 58.3)	22.9 (20.7, 25.1)	86.7 (85.2, 88.3)	57.3 (55.9, 58.8)
NDH or diabetes: FPG ≥6.1mmol/l	66.1 (60.1, 72.1)	55.2 (53.9, 56.6)	8.5 (6.9, 10.0)	96.3 (95.3, 97.2)	55.9 (54.5, 57.2)
NDH or diabetes: HbA1c ≥6.0%	68.3 (64.1, 72.5)	56.3 (55.0, 57.7)	14.2 (12.7, 15.7)	94.4 (93.5, 95.3)	57.5 (56.1, 58.8)

Values given as % (95% CI)

Figure 8.7 and Figure 8.8 show that the prevalence of every baseline outcome increased in each LPRS risk group.

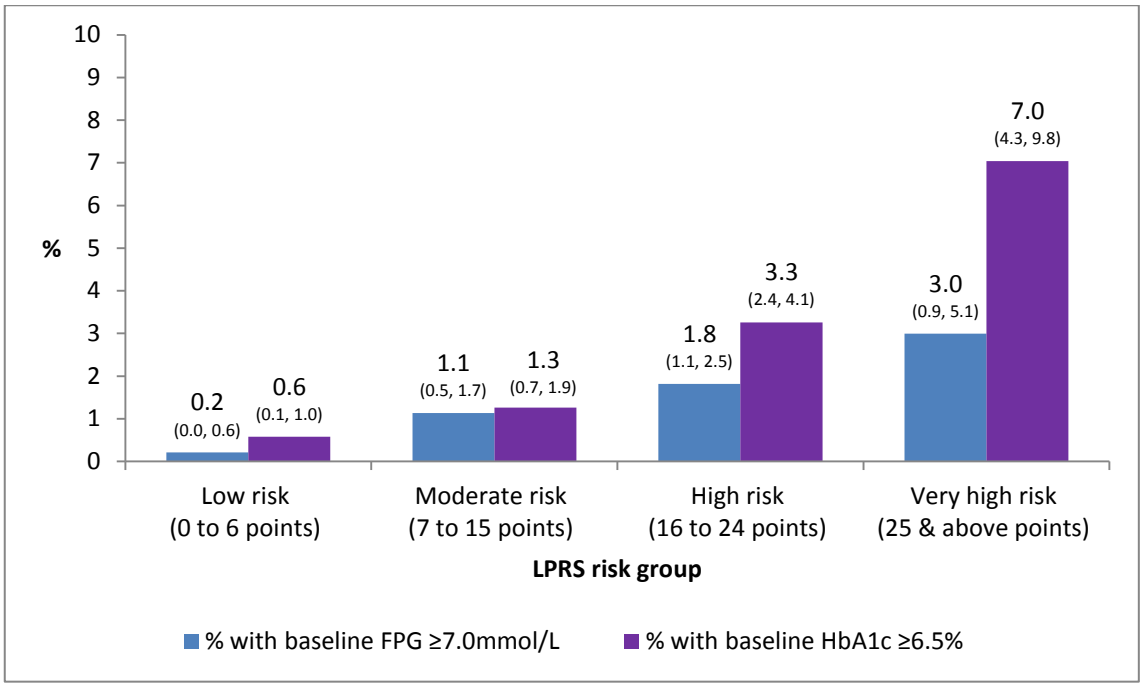


Figure 8.7 Percentage of individuals in each LPRS risk group with FPG ≥7.0mmol/l and HbA1c ≥6.5% at baseline

Data displayed as % (95% CI)

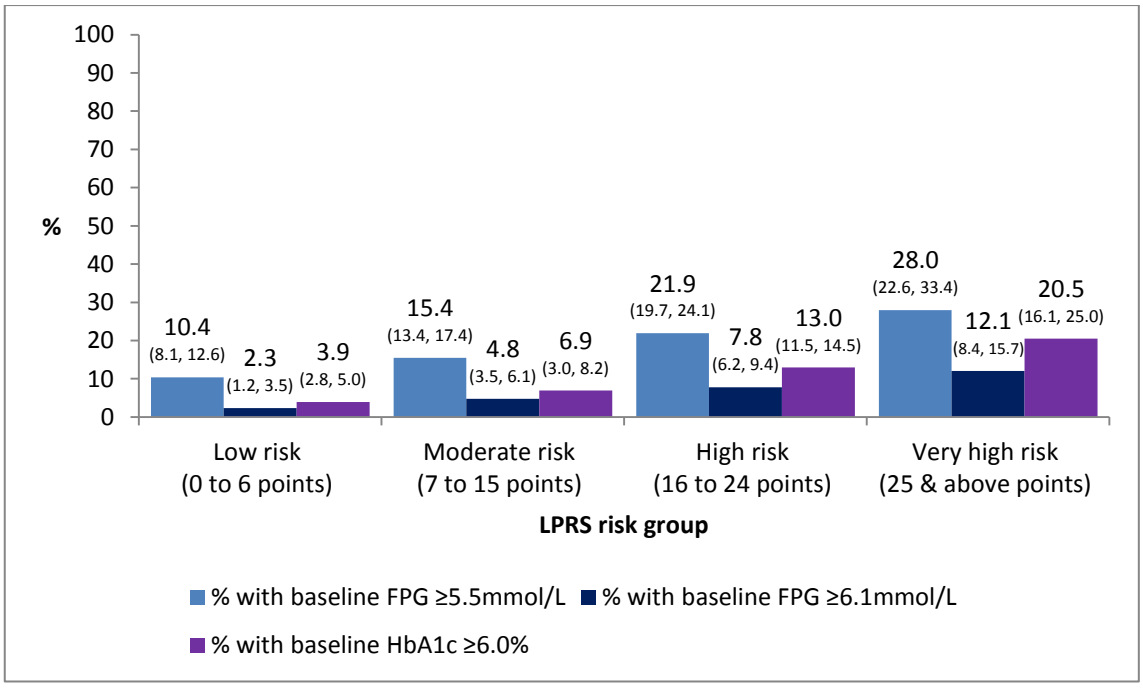


Figure 8.8 Percentage of individuals in each LPRS risk group with FPG ≥5.5mmol/l, FPG ≥6.1mmol/l and HbA1c ≥6.0% at baseline

Data displayed as % (95% CI)

8.4.5 LPRS for longitudinal outcomes

The AUROCs of the longitudinal outcomes ranged from 0.683 to 0.757, with the AUROCs for the HbA1c outcomes being higher than the AUROCs for the FPG outcomes. Most Brier scores were higher than their associated outcome index variances, with the exception of the HbA1c outcome at Wave 6 which had a Brier score marginally lower than its associated outcome index variance.

Table 8.12 Discrimination and calibration of LPRS risk score for various binary longitudinal diabetes outcomes in ELSA dataset

Outcome	**Prevalence**	**AUROC** (95% confidence interval)	**Brier score** (Outcome index variance)
Diabetes: FPG≥ 7.0mmol/l at Wave 4 or doctor diagnosed diabetes within four years	5.56% (4.86, 6.26)	0.706 (0.676, 0.737)	0.0650 (0.0525)
Diabetes: HbA1c ≥ 6.5% at Wave 4 or doctor diagnosed diabetes within four years	9.41% (8.52, 10.3)	0.704 (0.679, 0.730)	0.0869 (0.0852)
Diabetes: FPG≥ 7.0mmol/l at Wave 6 or doctor diagnosed diabetes within eight years	10.3% (9.02, 11.6)	0.683 (0.657, 0.709)	0.0931 (0.0923)
Diabetes: HbA1c ≥ 6.5% at Wave 6 or doctor diagnosed diabetes within eight years	13.3% (12.1, 14.6)	0.757 (0.727, 0.788)	0.0804 (0.0831)
Doctor Diagnosed within eight years	8.03% (7.25, 8.82)	0.701 (0.6735, 0.728)	0.0795 (0.0739)

Table 8.13 displays using the LPRS, with cut-point ≥0.155, gave very high NPV though it yielded low PPV for the various binary longitudinal outcomes. The cut-point produced a reasonable balance of sensitivity and specificity for the various outcomes.

Table 8.13 Sensitivity, specificity, PPV, NPV, proportion correctly classified and proportion classified as high risk of LPRS, with cut-point ≥0.155, for various binary longitudinal outcomes in ELSA dataset

Outcome	Sensitivity	Specificity	PPV	NPV	Correctly Classified
Diabetes: FPG≥ 7.0mmol/l at Wave 4 or doctor diagnosed diabetes within four years	74.0 (68.6, 79.3)	56.1 (54.7, 57.4)	9.0 (7.7, 10.3)	97.3 (96.7, 98.0)	55.5 (54.1, 56.9)
Diabetes: HbA1c ≥ 6.5% at Wave 4 or doctor diagnosed diabetes within four years	73.2 (69.0, 77.4)	57.0 (55.6, 58.4)	15.0 (13.4, 16.6)	95.3 (94.5, 96.2)	57.9 (56.6, 59.3)
Diabetes: FPG≥ 7.0mmol/l at Wave 6 or doctor diagnosed diabetes within eight years	69.9 (65.5, 74.2)	56.7 (55.3, 58.1)	15.6 (13.7, 17.5)	94.3 (93.0, 95.5)	58.1 (56.7, 59.4)
Diabetes: HbA1c ≥ 6.5% at Wave 6 or doctor diagnosed diabetes within eight years	69.8 (65.9, 73.6)	57.6 (56.2, 59.0)	20.2 (18.2, 22.2)	92.5 (91.3, 93.8)	59.2 (57.8, 60.6)
Doctor Diagnosed within eight years	73.0 (68.4, 77.6)	56.3 (55.0, 57.7)	12.7 (11.3, 14.2)	96.0 (95.2, 96.8)	57.7 (56.3, 59.0)

Values given as % (95% CI)

Figure 8.9 and Figure 8.10 display that the prevalence of every longitudinal outcome increased in each LPRS risk group.

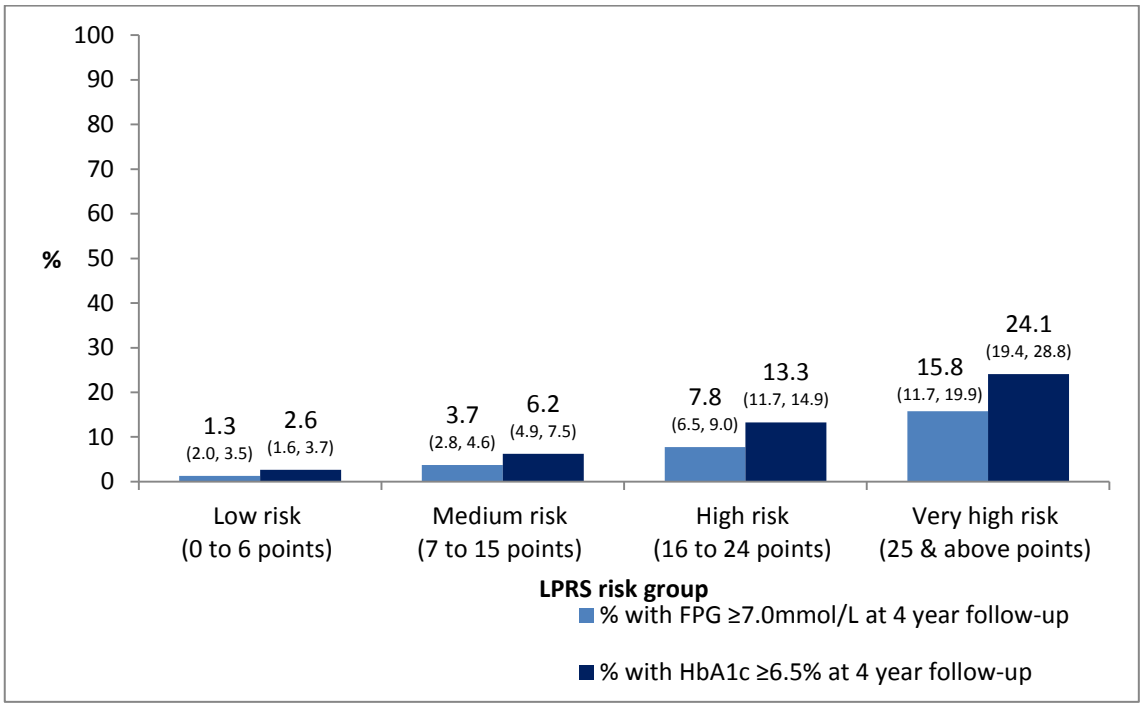


Figure 8.9 Percentage of individuals in each LPRS risk group with FPG ≥7.0mmol/l and HbA1c ≥6.5% at four year follow-up

Data displayed as % (95% CI)

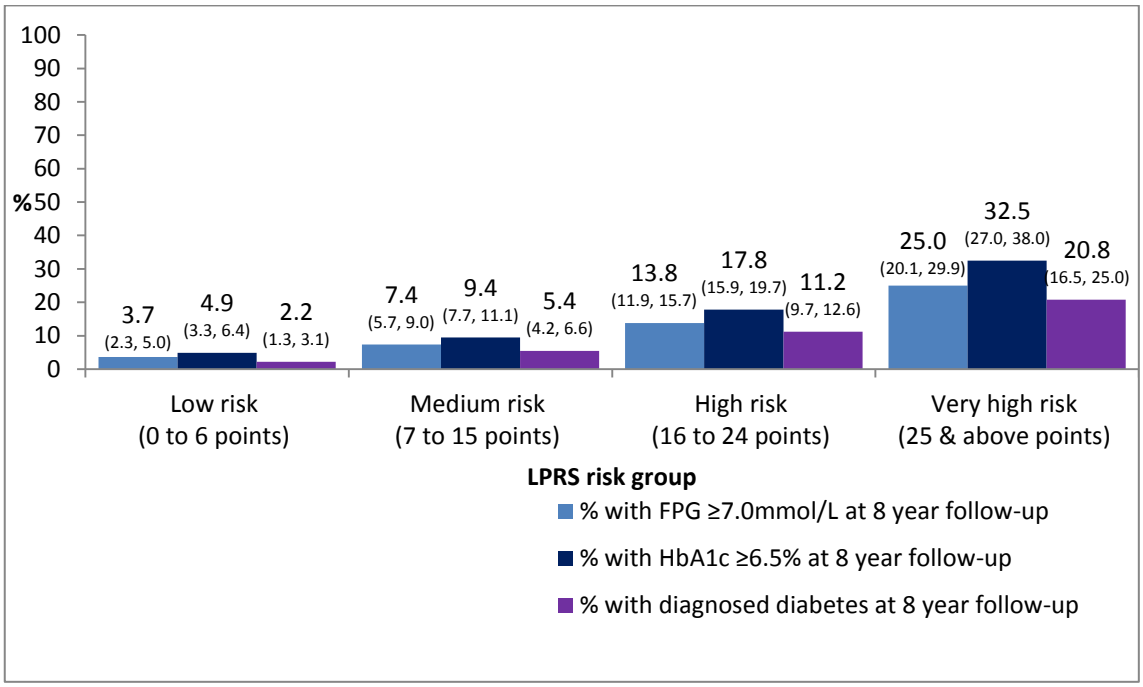


Figure 8.10 Percentage of individuals in each LPRS risk group with FPG ≥7.0mmol/l, HbA1c ≥6.5% and diagnosed diabetes at eight year follow-up

Data displayed as % (95% CI)

8.4.6 Two-staged approach with LPRS as first stage for longitudinal outcomes

Table 8.14 shows the two-stage screening programme with the LPRS as the first stage produced AUROCs ranging between 0.699 and 0.722 for the various blood tests and cut-points used in the second stage.

Table 8.14 Discrimination of two-stage screening programme, with LPRS as first stage and various blood tests as second stage, for binary longitudinal outcome of doctor diagnosed diabetes within 8-years in ELSA dataset

Blood screening test used	AUROC
FPG (with NDH defined as FPG≥5.5mmol/l)	0.718 (0.686, 0.787)
FPG (with NDH defined as FPG≥6.1mmol/l)	0.699 (0.668, 0.729)
HbA1c	0.722 (0.692, 0.752)

Table 8.15 shows that the specificity and NPV of the two-stage screening programme's decision were both high for the various blood tests and cut-point used in the second stage. Between 3.9% and 10.5% of individuals were classified as high risk depending on the blood test and cut-point used. Using HbA1c ≥6.0% as the cut-point in the second stage yielded the best combination of predictive diagnostics with sensitivity and PPV both being a little less than the highest produced.

Table 8.15 Predictive diagnostics of two-stage screening programme, with LPRS as first stage and various blood tests as second stage, for binary longitudinal outcome of doctor diagnosed diabetes within eight years in ELSA dataset

Blood test used	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Correctly Classified (%)	Classified as high risk (%)
FPG (with NDH defined as FPG≥5.5mmol/l)	44.0 (37.6, 50.4)	92.4 (91.5, 93.3)	33.6 (28.7, 38.5)	95.0 (94.2, 95.7)	88.5 (87.5, 89.5)	10.5 (9.5, 11.6)
FPG (with NDH defined as FPG≥6.1mmol/l)	26.7 (20.9, 32.6)	98.1 (97.6, 98.6)	55.1 (46.1, 64.0)	93.9 (93.1, 94.7)	92.4 (91.5, 93.2)	3.9 (3.2, 4.6)
HbA1c	39.7 (34.6, 44.8)	96.4 (95.8, 97.0)	48.7 (42.9, 54.5)	94.8 (94.1,95.5)	91.8 (91.0, 92.6)	6.5 (5.8, 7.2)

Values given as % (95% CI)

Figure 8.11 presents the prevalence of doctor diagnosed diabetes within eight years for the different risk groupings of two-staged baseline screening with LPRS as the first stage. The groupings which are below the cut-off for receiving an intensive lifestyle intervention (LPRS<0.155 and LPRS≥0.155 & blood glucose in the normal range) contained between 89.5% and 96.1% of individuals depending on the blood glucose test and cut-point used for the second stage of screening. These groupings had a low prevalence of the outcome, lower than observed for the whole dataset, with the exception of individuals with LPRS≥0.155 and FPG<6.1mmol/l which was marginally above the prevalence in the whole dataset.

The group of individuals with LPRS ≥0.155 and NDH has been combined with the group with LPRS ≥0.155 and a glucose measurement in the diabetes range, meaning the prevalence is shown for individuals with a positive screening as a whole. Around half the individuals screening positive using either HbA1c ≥6.0% or FPG ≥6.1mmol/l as the cut-point in the second stage were diagnosed with diabetes within eight years; this dropped to around a quarter when FPG ≥5.5mmol/l was used as the cut-point in the second stage.

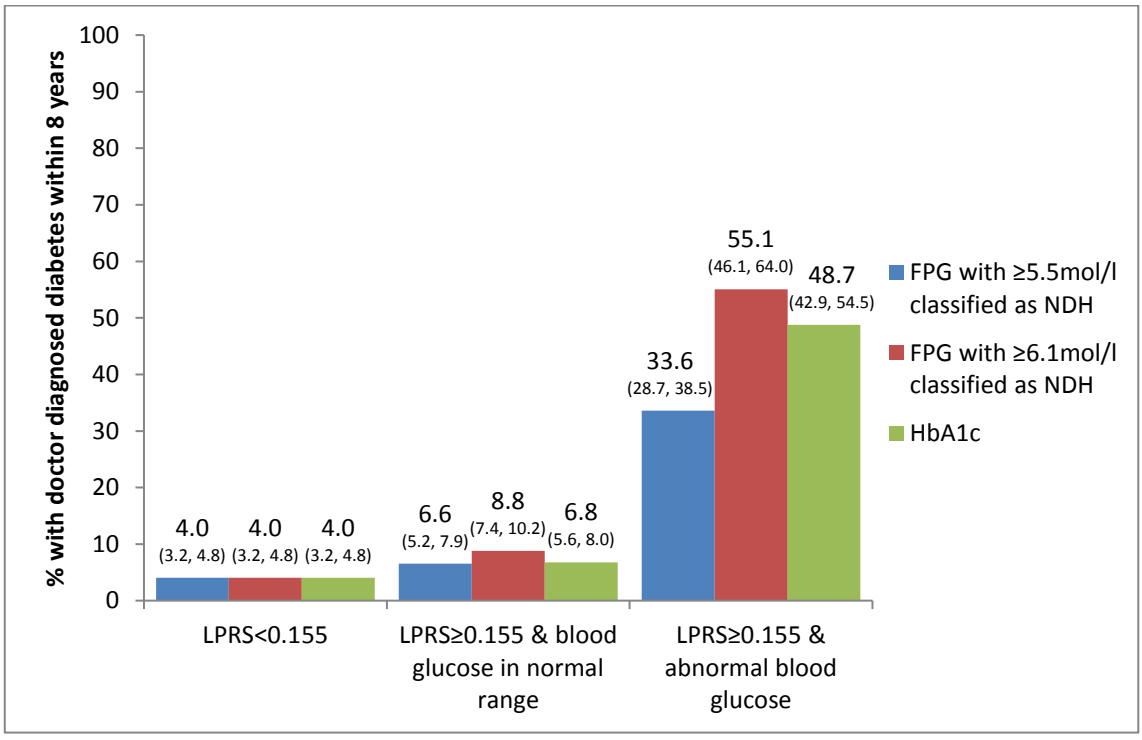


Figure 8.11 Percentage of individuals from different groupings of two-stage (with LPRS used at stage one) baseline screening being diagnosed with diabetes by a doctor within eight years shown for various baseline blood tests

Data displayed as % (95% CI)

8.5 Discussion

This chapter reports the first prospective external validation of both the LSA and LPRS using a nationally representative dataset. The LSA and LPRS discriminated well in this population for longitudinal diabetes outcomes, as well as for binary blood glucose outcomes at baseline. The recommended cut-points of the LSA, ≥16, and LPRS, ≥0.155, identified most individuals who were diagnosed with diabetes within eight years as being high risk; however many individuals who were not diagnosed with diabetes within eight years were also identified as high risk. This is acceptable though since the RATs are intended to be the first stage of a two-stage screening programme, thus they should be viewed as reducing the number of individuals requiring a blood test. Two-stage screening with either the LSA or LPRS as the first stage and a blood test as the second identified a small proportion of the population with a substantially increased risk of developing diabetes in the eight years which followed.

8.5.1 Dataset and variables

The dataset used was developed with the purpose of being nationally representative of individuals over the age of 50 years old. Although the proportion of women was higher than in England as a whole (195). However the underrepresentation of men was not pronounced enough to cause concern, with 44.5% of individuals included being male compared to estimated 48.9% of the population aged 50- 75 years old.

This dataset for the analyses was limited to 50-75 year olds as the RATs were designed for 40-75 year olds, and the dataset was developed to be a representative sample of people over 50 years old. Applying the RATs either alone or as part as a two-stage screening programme to 40-75 year olds, as recommended by the NICE guidelines, is likely to yield a lower proportion of the population categorised as being high risk, since age is a well-known T2DM risk factor and is included in both RATs (46,52). Furthermore, before being applied in age groups outside those of the range of this validation, an assessment of the performance of the RATs should be carried out using a dataset which includes the age group of interest.

One weakness of the data used was that individuals' family history of diabetes was not available at baseline, in Wave 2, and therefore parents' diabetes at final (eight year) follow-up was imputed for this variable. While this may not be perfect, given the age range of the participants, it was assumed that parents' history would be relatively stable over time (196). Yet the variable probably underestimated first degree family history of diabetes as it did not include siblings with diabetes, this may have caused the RATs to underperform slightly. This imputation is better than not using family history to calculate the score as was the case with the only existing geographical external validation for prevalent outcomes, since this will lead to the risk being underestimated for those with a family history of diabetes (25). That validation also had the weakness of including 16- 39 year olds, for whom the RATs were not developed and population wide use of RATs are not recommended.

Another issue was the missing data in three of the risk factors (body mass index (BMI), family history of diabetes and waist circumference) as well as all the outcome variables. The main analyses used multiple imputation to help overcome any bias caused by the missing data as well as the reduction in the sample size. Due to the high levels of the missing data, 50 imputations were used; as well as sensitivity analyses using complete data only for each analysis. The missing at random assumption is impossible to test using observed data and thus sensitivity analyses are a sensible check of the performance in the main analyses. Reassuringly, the RATs both performed better in the complete-case analyses than the analyses with multiple imputed data meaning the main analyses are more conservative.

Interestingly the proportion of individuals reporting antihypertensive use was lower in the ELSA dataset than in the Leicestershire based dataset, 12.5% compared to 23.8% and 22.6%. This may be because the variable was self-reported in ELSA rather than taken from medical records as should ideally be done but could also be reflective of the individuals in Leicestershire based datasets being at increased risk compared to the country as a whole. If the reason is the former this is likely to have caused underperformance of the LPRS in these analyses.

8.5.2 Analyses

The AUROC was chosen as the metric to assess discrimination since it is the commonly used measure in this field (55,57,59,99). As highlighted in Chapter 3, the discrimination is often of greatest concern to those developing or selecting a RAT or screening programme for use in practice; since the AUROC indicates the ability to correctly discern between individuals with and without the outcome of interest.

The Brier score was chosen to measure calibration. Though it should be noted that the Brier score measures overall fit of a model, of which calibration is a component rather than calibration alone (100,101). The Brier score was chosen since it is more consistent than tests of perfect calibration, such as the Hosmer-Lemeshow test (73). The Brier score is affected by the prevalence of the outcome, with a decrease in the prevalence leading to a decrease in the outcome index variance, which is the Brier score yielded from assigning each individual the prevalence as their prediction of the outcome (102). As the differently defined outcomes had differing prevalences of the outcome, the outcome index variance was displayed next to the Brier score in the tables to allow comparisons to be made to the non-informative model in each case.

A limitation of the analyses in this chapter was they did not include two of the RATs that are available in the United Kingdom (UK), namely the Cambridge Risk Score (CRS) and QDrisk (47,48). This was due to ELSA not containing variables of the RATs, in particular prescribed steroids for the CRS and the Townsend score or information to calculate it for QDrisk. A comparison of the performance of all the available RATs in the UK for both prevalent NDH or undiagnosed T2DM and incident T2DM would be helpful to those trying to select between them; especially since a study carried out by Gray et al. found differing proportions of the population being placed into the highest risk group for the different RATs (197).

8.5.3 LSA Results

Using the LSA alone to discriminate the cross-sectional binary outcomes yielded good levels of discrimination for identifying individuals currently in the diabetes range of the blood test measurements. The AUROCs were lower for identifying individuals in the abnormal range; this drop is similar to the one seen

for an adult population in the United States of America (USA) used to assess the performance of Heikes et al.'s Diabetes Risk Calculator (62,118). The AUROC of detecting abnormal FPG, the outcome which the RAT was developed using, was a little lower than the AUROC yielded on the population in which the LSA was developed, 0.67 compared to 0.69. This slight reduction may be due to the age range of dataset being smaller, 50- 75 year old compared to 40- 75 year olds. Interestingly the HbA1c defined outcomes had the highest AUROC for both the diabetes range outcomes and the abnormal range outcomes, being 0.73 and 0.68 respectively. These AUROCS are noticeably lower than the AUROC of 0.78 found for the outcome of prevalent NDH defined by HbA1c for individuals aged 16 years and older in the HSE dataset (25). This increased discrimination is likely due to the vastly wider age of that analysis.

The Brier scores were higher than their associated outcome index variances for the diabetes range outcomes indicating poor calibration. This is not surprising as the LSA was developed for the outcome of abnormal blood glucose hence the associated probabilities greatly overestimate the risk of having a blood glucose measurement in the diabetes range. Yet, disappointingly the Brier scores were a little higher than their corresponding outcome index variances for the abnormal blood glucose outcomes suggesting the calibration was worse than that of the non-informative model.

Using a cut-point of LSA ≥16 identified 63.3% of individuals as high risk and thus requiring a blood test. It missed only a few individuals with blood glucose in the diabetes range. The cut-point resulted in a large number of individuals requiring a blood test to identify one with blood glucose in the diabetes range, 53 individuals to yield one with FPG ≥7.0mmol/l and 31 to find one with HbA1c ≥6.5%. The recommended cut-point had satisfactory levels of sensitivity, around 80%, for all the NDH or undiagnosed T2DM outcomes. Furthermore, the extension of the blood glucose range of interest resulted in a significant drop in the number of individuals needing a blood test to identify one with the outcome of interest, with the numbers decreasing to between five and 13 depending on the blood test and cut-point used. Gray et al. found just under a quarter of individuals in their study to have a LSA ≥16 (197). One possible cause of this

lower proportion is due to the population being younger than the one included in this analysis, with an interquartile range of 44- 54 years. Hence if the score was applied to 40- 75 year olds the proportion of individuals identified as high risk and requiring a blood test would reduce from the 63.3% observed here.

The LSA had good levels of discrimination for the longitudinal binary diabetes outcome, with AUROC ranging from 0.70 to 0.73 for the various outcomes. On the other hand the Brier scores show the calibration was poor, though this is to be expected as the LSA was not developed for this outcome. Still the recommended cut-point of LSA ≥16 produced satisfactory levels of sensitivity for the various longitudinal diabetes outcomes. As expected, lower PPVs were seen for lower outcome prevalences (95). The PPV for the primary outcome is reasonable; it illustrates that one in ten individuals classified as high risk were diagnosed with diabetes by a doctor within eight years.

Good levels of discrimination for diagnosed diabetes by eight year follow-up were yielded from using the LSA as the first stage of a two-stage screening programme. Using HbA1c ≥6.0% as the cut-point in the second stage gave the best combination of predictive diagnostics with sensitivity being 47.0% and PPV being 47.3%. The PPV indicates around half of individuals with LSA ≥16 followed by HbA1c ≥6.0% were diagnosed with diabetes within eight years. Although high PPVs are seen for the two-stage baseline screening programme with the LSA as the first stage, the sensitivity levels highlight the importance of the reassessment of individuals' risk in three or five years time as recommended in NICE Public Health Guidelines 38 (28). One limitation of the analysis carried out in this chapter is that, due to the data available, assessing the use of the RATs and blood tests iterative, as suggested by NICE guidelines, was not feasible.

8.5.4 LPRS results

Using the LPRS alone for discriminating the cross-sectional binary outcomes produced good levels of discrimination for detecting individuals that were currently in the diabetes range of the blood test measurements, though the AUROC dropped for recognising individuals with abnormal blood glucose measurements. The AUROCS were lower for the FPG outcomes than those observed in the original external validation of the LPRS, 0.68 compared to 0.71 for undiagnosed T2DM and 0.65 rather than 0.69 for NDH or undiagnosed T2DM. However, the AUROC was higher for the outcome of HbA1c ≥6.5%, 0.72 compared to 0.69; and the same for HbA1c ≥6.0%, both 0.67. As with the LSA, the HbA1c defined outcomes had the highest AUROC for both the diabetes range outcomes and the NDH or undiagnosed T2DM outcomes. Once again the AUROCS were noticeably lower than the AUROC of 0.80 yielded for the outcome of prevalent NDH defined by HbA1c for individuals in the HSE dataset (25). Although, as stated earlier this increased discrimination is to be expected due to the significantly increased age range in that analysis. As expected, poor calibration was seen for the diabetes range outcomes. More significantly, the calibration of the LPRS was similar to that of the non-informative model for the abnormal blood glucose outcomes.

Selecting a cut-point of LPRS ≥0.155 identifies 46.0% of individuals as high risk and thus requiring a blood test. This cut-point has satisfactory levels of sensitivity, around 75% for the two diabetes range outcomes. However it led to vast numbers of individuals requiring a blood test to identify one with blood glucose in the diabetes range, 50 individuals to yield one with FPG ≥7.0mmol/l and 26 to find one with HbA1c ≥6.5%. The sensitivity dropped slightly when using this cut-point to identify individuals in the abnormal range outcomes. However, the numbers of individuals needing a blood test to identify one with the outcome of interest were considerably lower for the abnormal range outcomes, between four and 12 depending on the blood test and cut-point used.

The LPRS produced AUROCs ranging between 0.68 and 0.76 for the different longitudinal binary diabetes outcome which indicates good levels of discrimination. Though, as expected the Brier scores show the calibration was poor since the LPRS was not developed for these outcomes. The

recommended cut-point of LPRS ≥0.155 yielded reasonable levels of sensitivity for the various longitudinal diabetes outcomes. As was the case for the LSA, PPVs were positively correlated with outcome prevalences. The PPV for the primary outcome indicates that one in eight individuals classified as high risk were diagnosed with diabetes by a doctor within eight years.

Using the LPRS as the first stage of a two-stage screening programme gave good levels of discrimination. As was the case when using the LSA as the first stage, using HbA1c ≥6.0% as the cut-point in the second stage produced the best combination of predictive diagnostics. PPV indicates around half of individuals with LPRS ≥0.155 followed by HbA1c≥6.0% were diagnosed with diabetes within eight years. The sensitivity levels of the two-stage baseline screening programme with LPRS as the first stage, underline the need for reassessment of individuals' risk in future for those currently deemed not to require an intensive lifestyle intervention.

8.6 Conclusion and implications

Firstly, the LSA and LPRS have produced comparable performance in discriminating cross-sectional outcomes in the nationally representative dataset used in this chapter as they did for the Leicestershire based datasets in which they were originally developed and externally validated. Therefore they can be used across England for the purpose of identifying individuals who currently have NDH or undiagnosed T2DM.

Both RATs produced good levels of discrimination for identifying individuals who will progress to diabetes in the years that follow. Furthermore, implementing either RATs as the first stage of the two-stage baseline screening recommended by the NICE results in good discrimination and the identification of a small proportion of individuals with a markedly high risk of progressing to doctor diagnosed diabetes in the years that follow. Either RAT could be utilised as a useful tool to identify high risk people in the National Health Service (NHS) Diabetes Prevention Programme (DPP), which recently launched, since they have both been shown to be validated in identifying individuals who will develop diabetes unless they receive interventions.

Chapter 9: Discussion

9.1 Chapter outline

This chapter concludes the thesis with a general discussion of the work presented and of potential further research in the area. Strengths and limitations of the work in this thesis are also highlighted.

9.2 Summary of findings

Risk assessment tools (RATs) are advocated as one method for tackling the rise in prevalence of Type 2 diabetes mellitus (T2DM) (28,38). RATs help to optimise the resources required to detect individuals with non-diabetic hyperglycaemia (NDH) or undiagnosed T2DM. Additionally, they can make blood tests more acceptable to individuals than under population level screening, since they have been identified as being at an increased risk of the outcome compared to the population as a whole. This thesis includes work on the identification, development and validation of RATs for NDH and T2DM outcomes; informing those selecting or developing a RAT with such an outcome in particular. Additionally, work on methodological issues around RAT development is presented in this thesis, adding to the knowledge in the field of RATs as a whole.

Chapter 3 presented a systematic review of RATs which detect those at high risk of NDH in the general population. This was the first systematic review with a search strategy that focuses on finding RATs that screen for individuals with NDH, with existing systematic reviews in the field having found only three RATs for the outcome of prevalent NDH or undiagnosed T2DM (54-59). Eighteen RATs which detect individuals with prevalent NDH or with prevalent NDH or undiagnosed T2DM were summarised and critiqued, aiding those wishing to use such a RAT in their selection or development of an appropriate RAT. Additionally, the chapter emphasised the methodological issues highlighted by the previous systematic reviews in the area. Treatment of missing and continuous data were often not justified in papers detailing the development of RATs. Furthermore, external validations were often overlooked and impact studies were rare.

Chapter 4 detailed an empirical comparison of logistic regression, decision trees and SVMs for developing RATs for the outcome of prevalent impaired glucose regulation (IGR) or undiagnosed T2DM in a cross-sectional dataset. This was the first empirical comparison of methods for developing RATs for a medical outcome to include an external validation dataset (124-128). Despite the inclusion of extensions of the decision tree method such as bagging and

boosting, logistic regression and linear SVMs perform the best statistically in the external dataset. As linear SVMs do not provide a simple educational message in the way logistic regression can, this thesis suggests the use of logistic regression for developing RATs for prevalent NDH or undiagnosed T2DM. In addition, Chapter 4 assessed the effects of differing the sample size of the development dataset on the performance of each of the methods included in the empirical comparison through a resampling study. The results highlight the improvement in reliability and performance yielded by increasing the number of events per variable (EPV). A minimum of 20 EPV are recommended, with careful consideration of the evidence for including each candidate variable when the number of EPV is between 20 and 50.

Chapter 5 described the novel application of the chain event graph (CEG) method to develop a RAT for the outcome of prevalent NDH or undiagnosed T2DM. Issues with implementing CEGs for this novel application were overcome to allow the technique to be utilised to produce a RAT for the outcome of IGR or undiagnosed T2DM with good internal discrimination. However, the discrimination dropped noticeably in the external dataset and therefore the method should not be used over logistic regression or linear SVMs, which performed better externally.

Chapter 6 assessed whether the Leicester Self-Assessment (LSA) needed to be updated in light of the increased use of glycated haemoglobin A1c (HbA1c) as the blood test in practice. A RAT, which could be calculated using pen and paper, and a RAT which would require an electronic device to calculate were developed using logistic regression with an outcome of HbA1c ≥6.0%. The pen and paper RAT had similar discrimination to the LSA for detecting the outcome of HbA1c ≥6.0% in an external dataset. While the electronic RAT, which has practical disadvantages, performed only marginally better than the LSA for detecting HbA1c ≥6.0% in the external dataset. Consequently, neither RAT should replace the LSA in practice, as the LSA was not meaningfully outperformed by either updated RAT.

Chapter 7 established risk groups for the Leicester Practice Risk Score (LPRS) allowing consistent advice to be given across different general practices when

utilising the RAT. These risk groups are currently being incorporated by two providers into their database systems which are used by many general practices, namely Vision and SystmOne. Giving all individuals in the high or very high risk groups a blood test would identify around 75% of individuals with HbA1c ≥6.0%. However, this is likely to be unachievable for most general practices given the current resources available, even though it should be cost saving in the long term. A potentially viable screening strategy which will require less resources to implement is suggested. The strategy is to invite all individuals identified as very high risk, roughly 17% nationally, to targeted blood test appointments, then opportunistically offer blood tests to individuals identified as high risk when they visit the practice for another consultation.

Chapter 8 presented an external national validation of the LSA and LPRS using a nationally representative longitudinal dataset. The validation found the LSA and LPRS had comparable discrimination for prevalent NDH and prevalent NDH or undiagnosed T2DM in the nationally representative dataset as they did in the Leicestershire based datasets in which they were developed and validated. Consequently they can be used across England to identify individuals who currently have NDH or undiagnosed T2DM. In addition both RATs discriminated well the individuals who will progress to diabetes within eight years from those who will not, despite poor calibration. Importantly, implementing either risk score as the first stage of a two-stage baseline screening recommended by the National Institute for Health and Care Excellence (NICE) resulted in the identification of a small proportion of individuals with a decidedly increased risk of progressing to doctor diagnosed diabetes in the years that follow. Therefore the use of either RATs in practice across England is advocated, since both identify individuals who will develop diabetes unless they receive interventions when utilised in the recommended two-stage screening programme. The results in Chapter 8 also support the use of HbA1c as the screening test for individuals identified as high risk by a RAT rather than fasting plasma glucose (FPG).

9.3 Strengths and limitations

A key strength of the work carried out in this thesis is the consideration of the use in clinical practice of RATs throughout, alongside the statistical performance. This is of great importance as previous systematic reviews of RATs for detecting individuals at high risk of diabetes and the one carried out in the thesis have reported that although numerous RATs have been developed, very few are used in practice (54-59,99). One issue that arises when considering both the use in practice and statistical performance of self-assessment RATs is whether to categorise continuous variables. Doing so allows the RATs to be calculated by hand rather than requiring an electronic device, however studies have demonstrated that the practice damages the statistical performance (187). For these reasons, when developing RATs intended to be used for self-assessment logistic regression was utilised to create two RATs one with the continuous variables kept continuous and the other with categories for the continuous variables. In addition, when developing the RAT with categories, advice was followed on keeping the variables continuous during risk factor selection.

Chapters 4-6 in which RATs were developed and Chapter 7 in which risk groups were selected for the LPRS used an external dataset, Screening Those at Risk (STAR), to assess the validity of the RATs and groups developed. This was a strength to the empirical comparison as no previous empirical comparison of binary medical outcomes in cross-sectional data has included an external validation (124-128). A limitation of this work is the dataset used for external validation, although from a different study, was only a temporal validation rather than a geographical validation as is best practice (75,87). STAR had the same standard operating procedures as ADDITION-Leicester, it recruited individuals between 2002 and 2004 while ADDITION-Leicester recruited individuals between 2004 and 2009. Both recruited individuals from the same group of the population, 40- 75 years old or 25- 75 years old for individuals not of white European ethnicity, although STAR required individuals to have at least one recognised risk factor for T2DM while ADDITION-Leicester did not (46,88).

The external national validation of the LSA and LPRS for prevalent NDH and prevalent NDH or undiagnosed T2DM, overcame issues of previous external validation of the RATs. Namely, the STAR dataset being from the same geographical area as the development dataset; while the validation published by another group using the Health Survey for England dataset did not use family history, an important risk factor, in its calculations of the risk scores as well as including 16- 39 year olds for who the RATs were not developed and risk scores are not recommended for at a population level (25,28). Another strength of this validation was that it also assessed the performance of using the risk scores followed by blood tests at baseline in detecting those who prospectively develop diabetes. This is of upmost importance since this is the way in which RATs are intended to be used in practice.

The sample size of datasets used for external validation of RATs affects the reliability of the results, with studies suggesting datasets need to contain a minimum of 100 events and should preferably have more than 200 events (198,199). The STAR and English Longitudinal Study of Aging (ELSA) datasets used in the external validations in this thesis comfortably meet this benchmark for every outcome except for the outcome of undiagnosed T2DM in the ELSA dataset. Only around 100 individuals had this outcome when defining it by either FPG or HbA1c; however this was not the main outcome of interest with prevalent NDH or undiagnosed T2DM and incidence T2DM being of greater concern.

The decision to use an empirical comparison when evaluating the methods for developing RATs for the outcome of prevalent NDH or undiagnosed T2DM in Chapters 4 and 5 was taken due to the complexity of the causes of abnormal glucose. With the intricate nature of the condition meaning the assumptions required for a simulation study would be difficult to sensibly choose in order to reflect the real-world and may have unfairly favoured one of the methods assessed.

One major limitation of the work carried out in this thesis is the focus on developing RATs using a single dataset. Since RATs for NDH and undiagnosed T2DM have been shown to often be limited to populations very similar to those

in which they were developed, RATs developed using a single dataset may only be validated for a limited population (200,201). One solution to this issue that has been proposed is the use of meta-analysis of individual patient data (202). However, combining individual patient data using meta-analysis is a particularly challenging task; substantial barriers are faced even prior to the commencing of analysis, with considerable resources and willing collaborators both being required (203). The International Diabetes Federation's (IDF's) PREDICT-2 project is currently collecting appropriate datasets with the aim of developing a global RAT to predict future diabetes which can be adapted to specific countries (201). Fourteen of the 22 datasets obtained so far for this collaboration are cross-sectional and thus may contain the outcome of interest of this thesis, although the majority of individuals are Caucasian (204).

The treatment of missing data is a particularly complex issue, with the mechanism of missing being difficult to identify (136). When using the ADDITION-Leicester dataset for development in Chapters 4 and 6 multiple imputation of the candidate variables was carried out to avoid possible bias cause by the missing data. Fortunately, the levels of missing in the STAR dataset were low so the complete-case data was used to analyse the results in Chapters 4- 6, thus providing a check to the multiple imputation data also. Also the levels of missing for the variables required for Chapter 7 were low, 3.6%, so a complete case analysis was used. The resampling study in Chapter 4 used a complete case analysis only, due to computational intensity nature of the work carried out. Multiple imputation was also not possible in when developing radial SVMs in the empirical comparison in Chapter 4 or CEGs in Chapter 5, due to the computational complexity of the methods. A sensitivity analysis was included in Chapter 4 with a complete-case analysis using a reduced number of candidate variables, to check that none of the methods were detrimentally affected by the use of multiple imputation. High levels of missing data for the risk factors and outcomes included in the English Longitudinal Study of Aging (ELSA) analyses in Chapter 8 are of concern, however, with repeated blood tests in a longitudinal study this is unsurprising. The risk factors and outcomes were imputed using multiple imputation in an attempt to avoid bias in the analyses. However, whichever analysis is used it is likely some bias will have

resulted. A complete-case analysis was carried out as a sensitivity analysis, this reassuringly showed that the multiple imputation gave the most conservative estimates of performance.

Another limitation of the external prospective validation carried out in Chapter 8 was it did not include the Cambridge Risk Score (CRS) (48) or the QDrisk (47), which is also widely used in practice; due to some of the risk factors required not being included in the dataset.

9.4 Further research

The systematic review in Chapter 3 highlighted that many of the RATs for prevalent NDH or undiagnosed T2DM lack the external validations and impact studies which should be performed before they are implemented in practice. Therefore further research in the field should focus on addressing these shortcomings to make the RATs fit for purpose rather than development of even more RATs, particularly when studying populations for which RATs already exist.

As discussed in Section 9.3 there are methodological issues in applying the CEG method to develop RATs for cross-sectional outcomes which this thesis was unable to investigate sufficiently due to the time-consuming nature of the work, such as the order the risk factors are entered into the model or the way in which cut-points to group continuous variables are chosen. However, software is currently being developed to reduce the need for coding in developing CEGs which may make the investigation of these methodological issues viable in the next few years.

The field would benefit from an analysis of the impact of implementing the full two-stage screening programme recommended by NICE in 40- 75 year olds in the United Kingdom (UK) being carried out, rather than just the performance of baseline screening reported in Chapter 8. With individuals having their risk reassessed at the maximum time suggested by the programme for the category they are screened into, in addition to baseline two-stage screening. Calculating the total number of years of intervention offered and the number of years of intervention offered to those who go on to develop diabetes without interventions would be extremely informative to decision makers. Furthermore, this information could inform a cost effective analysis or impact study of using the two-stage screening programme followed by offering interventions where recommended in preventing and delaying T2DM. Ideally such an analysis would compare all RATs used in practice in the UK to assess diabetes risk (24,45-47). This analysis would require a dataset with each of the risk factors of RATs included and blood test measurements recorded every year, as well as the dates of any diagnosis of T2DM.

Finally, advancements in the area of joint modelling of longitudinal and survival data provide an opportunity for models of incident T2DM which take into account the dynamic risk change over time as a result in the change in different risk factors to be established (205,206). Such models would be helpful to inform and motivate those wishing to change their or other individuals' risk of developing the condition.

9.5 Final conclusions

RATs for identifying individuals at high risk of T2DM are recommended to tackle the rise of the disease. Increasing numbers of RATs have been developed to meet this demand, although how they will be applied in practice is often overlooked. Many RATs lack the external validations and impact studies required to support implementation in practice. It is hoped that the work carried out in this thesis informs those wishing to use or develop such a RAT, as well as having overcome some issues with implementing the LSA and LPRS in practice across England.

Appendices. Supplementary material

Appendix A: Supplementary material relating to systematic review

Appendix B: Supplementary material relating to Leicester Practice Risk Score groups

Appendix C: Supplementary material relating to English Longitudinal Study of Aging sensitivity analyses

Appendix A: Supplementary material relating to systematic review

The following terms and strategy, decided after discussing with a librarian, were used to search for articles on OvidSP for both the Medline search and the Embase search; only terms 17 and 31 were different as the Mesh terms in the two databases differs in these cases.

1. ((screen$ or diagnostic or score) adj5 risk).tw.
2. Predict$.ti.
3. (Predict$ adj5 (Outcome$ or Risk$ or Model$)).tw.
4. ((Variable$ or Criteria or Scor$ or Characteristic$ or Factor$) adj5 (Predict$ or Model$ or Decision$ or Identif$ or Prognos$)).tw.
5. Decision$.tw. adj5 ((Model$ or Clinical$).tw. or Logistic Models/)
6. (Prognostic adj5 (History or Variable$ or Criteria or Scor$ or Characteristic$ or Finding$ or Factor$ or Model$)).tw.
7. Predict$ adj5 model$.tw.
8. Predict$ adj5 equation.tw.
9. Predict$ adj5 rule.tw.
10. (risk adj5 (calculator or model or assessment)).tw.
11. algorithm.tw.
12. "recursive partition$".tw.
13. multivariate.tw.
14. Statistic$ adj5 model.tw.
15. Prediab$.ti.
16. Pre diab$.ti.
17. exp Prediabetic state/ (Medline) exp impaired glucose tolerance/ (Embase)
18. (glucose adj2 impair$).ti.
19. (glucose adj2 intol*).ti.
20. IGT.ti.
21. IFG.ti.
22. (impair$ adj2 glycem$).ti.
23. (impair$ adj2 glycaem$).ti.
24. (insulin adj2 resistan$).ti.
25. Type 2 Diab$.ti.
26. exp Insulin Resistance/
27. Type II Diab$.ti.
28. NIDDM.ti.
29. Non insulin dependent Diabetes.ti.
30. T2DM.ti.
31. exp Diabetes Mellitus, Type 2/ (Medline) exp non insulin dependent diabetes mellitus/ (Embase)
32. IGR.ti.
33. Impaired glucose regulation.ti.
34. Impaired glucose tolerance.ti.
35. impaired fasting glucose.ti.
36. glucose intolerance.ti.
37. obese diabetes.ti.
38. obesity diabetes.ti.
39. ((adult or mature or late) and onset).ti.
40. MODY.ti.
41. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14
42. 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40
43. 41 and 42
44. non-diabet$.ti.
45. 43 NOT 44
46. limit 45 to humans
47. limit 46 to English language

Figure 10.1 Search strategy of systematic review

General information
Title of paper
First author
Name of risk score
Difference to risk score in same paper (if appropriate)
Journal
Date published
Notes:

Internal sample information
Name of study which the data is from
Year
Country/countries
Primary reason for cohort
Sampling frame (inclusion/exclusion criteria and key characteristics)
Sample size
Outcome & definition used (if appropriate)
Outcome Incidence rate and (number of events) (if appropriate)
Notes:
Model information
Method for developing risk score
Number of variables considered for inclusion in the model
Variables included in final model
Variables considered but not included in final model
How model was selected? (variable selection)
Treatment of continuous data
Treatment of missing data
Method of choosing cut-off (was this decided in advance?)
Internal sensitivity/specificity at recommend cut-off (if reported)
Internal PPV/NPV at recommend cut-off (if reported)
% Needing further testing (if reported)
Area under ROC
Any internal validation of the model
Calibration
Way in which the risk score can be completed (e.g. self-assessment, G.P assessment etc.)
Notes:
External validation
Was an external validation carried out by the author in the same paper?
Year
Country/countries
Primary reason for cohort
Sampling frame
Sample size & number of events
Was the same outcome definition used?
Sensitivity/specificity at recommend cut-off (if reported)
PPV/NPV at recommend cut-off (if reported)
Area under ROC
Calibration
Any other external assessments of the model using this dataset
Notes:
Author's opinion
Intended use of risk score (who will use score and mechanism of implementation)
Recommended action should be taken by those who score above the cut-off
Strengths of score
Weaknesses of score
Notes:

Figure 10.2 Systematic review data extraction form

	Barriga[1] [46]	Barriga[2] [46]	DuBose [25]	Gray[1] [43]	Gray[2] [31]	Handlos[1] [32]	Handlos[2] [32]	Handlos[3] [32]	Handlos[4] [32]	Heikes [30]	Hische[1] [47]	Hische[2] [47]	Koopman [27]	Nelson [28]	Robinson[1] [26]	Robinson[2] [26]	Yu [23]	Xin [24]
Age	√	√	√	√	√	√	√	√	√	√	√		√	√	√	√	√	√
BMI[a]	√	√	√	√	√	√	√	√	√				√		√	√	√	
Cholesterol medication						√	√	√	√									
CVD[a]						√	√		√									
Education						√	√		√						√	√		
Ethnicity				√	√	√	√		√					√	√	√	√	
Family history of T2DM[a]				√	√	√	√	√	√	√			√		√	√	√	√
Fast food intake						√	√		√									
Fasting glucose	√											√		√				
Fizzy drinks intake						√	√											
Fruit/vegetable intake						√		√							√	√		
Gender			√	√	√	√	√		√				√		√	√	√	
Gestational diabetes						√	√	√										
Height																	√	
History of high blood glucose															√	√		
Hypertension[a]				√	√	√	√	√	√				√	√	√	√	√	√
Macrosomia															√	√		
People per room in house						√												
Physical activity						√	√								√	√	√	
Resting heart rate			√										√					
Self-assessed health						√	√	√	√									
Systolic blood pressure											√	√						
Triglyceride														√				
Waist				√						√					√	√	√	√
Waist-to-hip ratio																		√
Weight[a]							√	√	√								√	

BMI – Body Mass Index, CVD – cardiovascular disease, T2DM – Type 2 diabetes Mellitus. Tools appear in same order as in other Tables so Barriga[1] is the tool by Barriga et al. given first in the other tables.

[a] Several variations on variable and tools ticked for this variable may include more than one of these variations, for example Handlos[1] includes both history of CVD and treatment for CVD.

Figure 10.3 Risk factors in each risk assessment tool identified in the systematic review

Appendix B: Supplementary materials relating to Leicester Practice Risk Score groups

Table 10.1 Sensitivity, specificity, PPV and NPV of the Initial risk groups' cut-points for outcome of FPG ≥5.5mmol/l in the internal and external datasets

Cut-point	ADDITION-Leicester				STAR			
	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
≥0.075	96.4 (95.3, 97.3)	12.1 (11.2, 13.1)	26.2 (25.1, 27.4)	91.2 (88.7, 93.3)	97.6 (96.4, 98.5)	5.0 (4.1, 5.9)	29.6 (28.0, 31.3)	83.6 (76.2, 89.4)
≥0.165	69.2 (66.8, 71.6)	52.5 (51.0, 54.0)	32.1 (30.5, 33.7)	84.0 (82.7, 85.4)	81.3 (78.6, 83.8)	37.2 (35.2, 39.2)	34.6 (32.6, 36.7)	82.9 (80.5, 85.2)
≥0.325	25.4 (23.2, 27.7)	89.4 (88.5, 90.3)	43.7 (40.4, 47.1)	78.7 (77.6, 79.8)	35.2 (32.2, 38.4)	81.2 (79.4, 82.7)	43.4 (39.8, 47.0)	75.4 (73.6, 77.1)

Values given are % (95% CI).

Table 10.2 Sensitivity, specificity, PPV and NPV of the Simplified risk groups' cut-points for outcome of FPG ≥5.5mmol/l in the internal and external datasets

Cut-point	ADDITION-Leicester				STAR			
	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
≥0.105	89.9 (88.2, 91.3)	25.6 (24.4, 26.9)	28.1 (24.4, 26.9)	88.6 (86.8, 90.3)	94.8 (93.2, 96.1)	13.6 (12.2, 15.1)	31.0 (29.3, 32.7)	86.5 (82.5, 89.9)
≥0.155	72.4 (70.1, 74.7)	48.7 (47.2, 50.1)	31.4 (29.8, 33.0)	84.5 (83.1, 85.9)	84.2 (81.7, 86.5)	33.5 (31.6, 35.5)	34.1 (32.2, 36.1)	83.8 (81.3, 86.2)
≥0.305	28.9 (26.6, 31.3)	86.9 (85.9, 87.9)	41.7 (38.7, 44.7)	79.0 (77.9, 80.2)	41.2 (38.0, 44.4)	77.8 (76.0, 79.5)	43.1 (39.9, 46.5)	76.4 (74.6, 78.1)

Values given are % (95% CI).

Appendix C: Supplementary materials relating to English Longitudinal Study of Aging sensitivity analyses

Table 10.3 Number of individuals with complete data for each analysis

	LSA	LPRS	LSA followed by HbA1c	LPRS followed by HbA1c	LSA followed by FPG	LPRS followed by FPG
Baseline HbA1c ≥6.0%	3,148	3,168	N/A	N/A	N/A	N/A
Baseline HbA1c ≥6.5%	3,148	3,168	N/A	N/A	N/A	N/A
Baseline FPG ≥5.5mol/L	2,272	2,282	N/A	N/A	N/A	N/A
Baseline FPG ≥6.1mol/L	2,272	2,282	N/A	N/A	N/A	N/A
Baseline FPG ≥7.0mol/L	2,272	2,282	N/A	N/A	N/A	N/A
HbA1c ≥6.5% at four year follow-up	2,680	2,680	N/A	N/A	N/A	N/A
FPG ≥7.0mol/L at four year follow-up	1,807	1,807	N/A	N/A	N/A	N/A
HbA1c ≥6.5% at eight year follow-up	2,677	2,700	N/A	N/A	N/A	N/A
FPG ≥7.0mol/L at eight year follow-up	1,577	1,585	N/A	N/A	N/A	N/A
Doctor diagnosed diabetes within eight years	3,883	3,916	3,135	3,155	2,263	2,273

Table 10.4 Discrimination and calibration of LSA risk score for various binary cross-sectional outcomes in ELSA dataset using complete-case analysis

Outcome	Prevalence	AUROC (95% confidence interval)	Brier score (outcome variance index)
Diabetes: FPG ≥ 7.0mmol/l (n= 2272)	1.14%	0.716 (0.640, 0.793)	0.0605 (0.0113)
Diabetes: HbA1c ≥ 6.5% (n=3148)	1.81%	0.762 (0.703, 0.821)	0.0614 (0.0178)
NDH or diabetes: FPG ≥ 5.5mmol/l (n= 2272)	15.0%	0.632 (0.601, 0.663)	0.130 (0.127)
NDH or diabetes: FPG ≥ 6.1mmol/l (n= 2272)	4.40%	0.688 (0.637, 0.739)	0.0746 (0.0421)
NDH or diabetes: HbA1c ≥ 6.0% (n=3148)	7.72%	0.696 (0.661, 0.730)	0.0886 (0.0712)

Table 10.5 Discrimination and calibration of LSA for various binary longitudinal diabetes outcomes in ELSA dataset using complete-case analysis

Outcome	Prevalence	AUROC (95% confidence interval)	Brier score (Outcome variance index)
Diabetes: FPG≥ 7.0mmol/l at Wave 4 or doctor diagnosed diabetes within four years (n=1807)	2.66%	0.734 (0.670, 0.797)	0.0629 (0.0259)
Diabetes: HbA1c ≥ 6.5% at Wave 4 or doctor diagnosed diabetes within four years (n=2680)	6.90%	0.742 (0.708, 0.775)	0.0821 (0.0643)
Diabetes: FPG≥ 7.0mmol/l at Wave 6 or doctor diagnosed diabetes within eight years (n=1577)	5.20%	0.701 (0.642, 0.761)	0.0712 (0.0493)
Diabetes: HbA1c ≥ 6.5% at Wave 6 or doctor diagnosed diabetes within eight years (n=2677)	8.97%	0.754 (0.724, 0.783)	0.0893 (0.0816)
Doctor Diagnosed within eight years (n=3883)	6.80%	0.740 (0.711, 0.768)	0.0839 (0.0634)

Table 10.6 Discrimination of two-stage screening programme, with LSA as first stage and various blood tests as second stage, for binary longitudinal outcome of doctor diagnosed diabetes within eight years in ELSA dataset using complete-case analysis

Blood screening test used	Prevalence	AUROC (95% confidence interval)
FPG (with NDH defined as FPG≥5.5mmol/l) (n=2263)	5.61%	0.750 (0.705, 0.794)
FPG (with NDH defined as FPG≥6.1mmol/l) (n=2263)	5.61%	0.731 (0.689, 0.774)
HbA1c (n=3135)	6.19%	0.774 (0.738, 0.810)

Table 10.7 Discrimination and calibration of LPRS score for various binary cross-sectional outcomes in ELSA dataset using complete-case analysis

Outcome	Prevalence	AUROC (95% confidence interval)	Brier score (Outcome variance index)
Diabetes: FPG ≥ 7.0mmol/l (n= 2282)	1.14%	0.706 (0.627, 0.785)	0.0392 (0.0113)
Diabetes: HbA1c ≥ 6.5% (n=3168)	1.80%	0.762 (0.699, 0.824)	0.0415 (0.0177)
NDH or diabetes: FPG ≥ 5.5mmol/l (n= 2282)	15.0%	0.623 (0.591, 0.654)	0.125 (0.127)
NDH or diabetes: FPG ≥ 6.1mmol/l (n= 2282)	4.38%	0.675 (0.624, 0.727)	0.0578 (0.0419)
NDH or diabetes: HbA1c ≥ 6.0% (n=3168)	7.77%	0.681 (0.647, 0.715)	0.0773 (0.0716)

Table 10.8 Discrimination and calibration of LPRS risk score for various binary longitudinal diabetes outcomes in ELSA dataset using complete-case analysis

Outcome	Prevalence	AUROC (95% confidence interval)	Brier score (Outcome variance index)
Diabetes: FPG≥ 7.0mmol/l at Wave 4 or doctor diagnosed diabetes within four years (n=1807)	2.69%	0.693 (0.620, 0.766)	0.0460 (0.0262)
Diabetes: HbA1c ≥ 6.5% at Wave 4 or doctor diagnosed diabetes within four years (n=2699)	7.00%	0.728 (0.693, 0.763)	0.0708 (0.0651)
Diabetes: FPG≥ 7.0mmol/l at Wave 6 or doctor diagnosed diabetes within eight years (n=1585)	5.24%	0.699 (0.641, 0.757)	0.0588 (0.496)
Diabetes: HbA1c ≥ 6.5% at Wave 6 or doctor diagnosed diabetes within eight years (n=2700)	9.15%	0.757 (0.727, 0.788)	0.0804 (0.0831)
Doctor Diagnosed within eight years (n=3916)	6.95%	0.737 (0.708, 0.765)	0.0709 (0.0646)

Table 10.9 Discrimination of two-stage screening programme, with LPRS as first stage and various blood tests as second stage, for binary longitudinal outcome of doctor diagnosed diabetes within 8-years in ELSA dataset using complete-case analysis

Blood screening test used	**Prevalence**	**AUROC** (95% confidence interval)
FPG (with NDH defined as FPG≥5.5mmol/l) (n=2273)	5.63%	0.739 (0.692, 0.787)
FPG (with NDH defined as FPG≥6.1mmol/l) (n=2273)	5.63%	0.728 (0.682, 0.774)
HbA1c (n=3155)	6.31%	0.738 (0.698, 0.777)

List of references

(1) International Diabetes Federation. IDF Diabetes Atlas. 7th ed. Brussels: International Diabetes Federation; 2015.

(2) Boyle J, Engelgau M, Thompson T, Goldschmid M, Beckles G, Timberlake D, et al. Estimating prevalence of type 1 and type 2 diabetes in a population of African Americans with diabetes mellitus. American Journal of Epidemiology 1999;149:55-63.

(3) Bruno G, Runzo C, Cavallo-Perin P, Merletti F, Rivetti M, Pinach S, et al. Incidence of type 1 and type 2 diabetes in adults aged 30-49 years: the population-based registry in the province of Turin, Italy. Diabetes care 2005;28:2613-2619.

(4) Holman N, Young B, Gadsby R. Current prevalence of type 1 and type 2 diabetes in adults and children in the UK. Diabetic Medicine 2015;32:1119-1120.

(5) Fowler M. Microvascular and Macrovascular Complications of Diabetes. Clinical Diabetes 2008;26(2):77-82.

(6) Melmed S, Polonsky K, Reed Larsen P, Kronenburg H. William's Textbook of Endocrinology. 12th ed.: Philadelphia: Elsevier/Saunders; 2011. p. 1371-1435.

(7) Hex N, Bartlett C, Wright D, Taylor M, Varley D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. Diabetic Medicine 2012;29(7):855-862.

(8) American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes 2012;35:64-71.

(9) Nathan D, Turgeon H, Regan S. Relationship between glycated haemoglobin levels and mean glucose levels over time. Diabetologia 2007;50:2239-2244.

(10) Stratton I, Adler A, Neil H, Matthews D, Manley S, Cull C, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000;321:405-412.

(11) Coutinho M, Gerstein H, Wang Y, Yusuf S. The relationship between glucose and incident cardiovascular events: a metaregression analysis of published data from 20 studies of 95,783 individuals followed for 12.4 years. Diabetes Care 1999;22:233-240.

(12) World Health Organization. Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus. 2011; Available at: http://www.who.int/diabetes/publications/report-hba1c_2011.pdf. Accessed 10/08, 2013.

(13) World Health Organisation. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. 2006; Available at: http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf Accessed 10/08, 2013.

(14) The International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. Diabetes Care 2009;32:1327-1334.

(15) Pyorala K, Pedersen TR, Kjekshus J, Faergeman O, Olsson AG, Thorgeirsson G. Cholesterol lowering with simvastatin improves prognosis of diabetic patients with coronary heart disease. A subgroup analysis of the Scandinavian Simvastatin Survival Study. Diabetes Care 1997;20:614-620.

(16) Heart Outcomes Prevention Evaluation Study Investigators. Effects of Ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus: results of the HOPE study and MICRO-HOPE substudy. Lancet 2000;355:253-259.

(17) UK Prospective Diabetes Study Group. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. BMJ 1999;317:703-713.

(18) Ray KK, Seshasai SR, Wijesuriya S, et al. Effect of intensive control of glucose on cardiovascular outcomes and death in patients with diabetes mellitus: a meta-analysis of randomised controlled trials. Lancet 2009;373:1765-1772.

(19) Porta M, Curletto G, Cipullo D, Rigault de la Longrais, R., Trento M, Passera P, et al. Estimating the Delay Between Onset and Diagnosis of Type 2 Diabetes From the Time Course of Retinopathy Prevalence. Diabetes Care 2014;37:1668-1674.

(20) Harris MI, Klein R, Welborn TA, Knuiman MW. Onset of NIDDM Occurs at Least 4-7 Yr Before Clinical Diagnosis. Diabetes Care 1992;15(7):815-819.

(21) International Diabetes Federation. Undiagnosed diabetes. 2013; Available at: http://www.idf.org/diabetesatlas/5e/undiagnosed-diabetes. Accessed 08/30, 2013.

(22) Khunti K, Davies M. Should we screen for type 2 diabetes: Yes. BMJ 2012;345(e4514).

(23) Griffin S, Borch-Johnsen k, Davies M, Khunti K, Rutten G, Sandbaek A, et al. Effect of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with type 2 diabetes detected by screening (ADDITION-Europe): a cluster-randomised trial. Lancet 2011;378(9786):156-167.

(24) Simmons R, Echouffo-Tchugui J, Sharp S, Sargeant L, Williams K, Prevost T, et al. Screening for type 2 diabetes and population mortality over 10 years (ADDITION-Cambridge): a cluster-randomised controlled trial. Lancet 2012;380(9855):1741-1748.

(25) NCVIN. NHS Diabetes Prevention Programme: Non-diabetic hyperglycaemia. Public Health England 2015;2015206.

(26) Gerstein H, Santaguida P, Raina P, Morrison K, Balion C, Hunt D. Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: A systematic overview and meta-analysis of prospective studies. Diabetes Research & Clinical Practice 2007;78(3):305-312.

(27) Morris D, Khunti K, Achana F, Srinivasan B, Gray L, Davies M, et al. Progression rates from HbA1c 6.0-6.4% and other prediabetes definitions to type 2 diabetes: a meta-analysis. Diabetologia. 2013;56(7):1489-1493.

(28) National Institute for Health and Care Excellence. Preventing type 2 diabetes: risk identification and interventions for individuals at high risk 2012; Available at: http://publications.nice.org.uk/preventing-type-2-diabetes-risk-identification-and-interventions-for-individuals-at-high-risk-ph38/recommendations. Accessed 08/06, 2013.

(29) American Diabetes Association. Summary of Revisions for the 2010 Clinical Practice Recommendations. Diabetes Care 2010;33.

(30) Tarasova V, Caballero J, Turner P, Inzucchi S. Speaking to Patients About Diabetes Risk: Is Terminology Important? Clinical Diabetes 2014;32(2):90-95.

(31) Tuomilehto J, Lindström J, Eriksson J, Valle T, Hämäläinen H, Ilanne-Parikka P. Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance. New England Journal of Medicine 2001;344(18):1343-1350.

(32) Knowler W, Barrett-Connor E, Fowler S, Hamman R, Lachin J, Walker E, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. New England Journal of Medicine. 2002;346(6):393-403.

(33) Gillies C, Abrams K, Lambert P, Cooper N, Sutton A, Hsu T, et al. Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis BMJ 2007;334:299.

(34) Li G, Zhang P, Wang J, Gregg E, Yang W, Gong Q, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. Lancet 2008;371:1783-1789.

(35) Gong Q, Gregg E, Wang Y, An Y, Zhang P, Yang W, et al. Long-term effects of a randomised trial of a 6-year lifestyle intervention in impaired glucose tolerance on diabetes-related microvascular complications: the China Da Qing Diabetes Prevention Outcome Study. Diabetologia 2011;54:300-307.

(36) NHS England. NHS Diabetes Prevention Programme (NHS DPP). 2016; Available at: https://www.england.nhs.uk/ourwork/qual-clin-lead/diabetes-prevention/. Accessed 07/23, 2016.

(37) UK National Screening Committee. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme. 2013; Available at: http://www.screening.nhs.uk/criteria. Accessed 03/28, 2014.

(38) Waugh N, Shyangdan D, Taylor-Phillips S. Screening for type 2 diabetes: a short report for the National Screening Committee. Health Technology Assessment 2013;17(35).

(39) Mostafa S, Davies M, Srinivasan B, Carey M, Webb D, Khunti K. Should glycated haemoglobin (HbA1c) be used to detect people with type 2 diabetes mellitus and impaired glucose regulation? Postgraduate Medical Journal 2010;86:656-662.

(40) Steyerberg E, Moons K, Van der Windt D, Hayden J, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Medicine 2013;10(2).

(41) Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd ed. New York, NY: John Wiley & Sons; 2000.

(42) Lindström J, Tuomilehto J. The diabetes risk score. Diabetes care 2003;26:725-731.

(43) Ryden L, Grant P, Anker S, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. European Heart Journal 2013;34(39):3035-3087.

(44) Collins G, Reitsma J, Altman D, Moons K. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Annals of Internal Medicine 2015;162(1):55-63.

(45) Gray LJ, Davies MJ, Hiles S, Taub NA, Webb DR, Srinivasan BT, et al. Detection of impaired glucose regulation and/or type 2 diabetes mellitus, using primary care electronic data, in a multiethnic UK community setting. Diabetologia 2012;55(4):959-966.

(46) Gray L, Tringham J, Davies M, Webb DR, Jarvis J, Skinner T, et al. Screening for type 2 diabetes in a multiethnic setting using known risk factors to identify those at high risk: a cross-sectional study. Vascular Health and Risk Management. 2010;6:837-842.

(47) Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. BMJ 2009;338(b880).

(48) Griffin S, Little P, Hales C, Kinmonth A, Wareham N. Diabetes risk score: towards earlier detection of Type 2 diabetes in general practice. Diabetes/Metabolism Research and Reviews 2000;16:164-171.

(49) Glumer C, Vistisen D, Borch-Johnsen K, Colagiuri S. Risk Scores for Type 2 Diabetes Can Be Applied in Some Populations but Not All. Diabetes Care 2006;29:410-414.

(50) Spijkerman A, Dekker J, Yuyan M, Nijpels G, Griffin S, Wareham N. The Performance of a Risk Score as a Screening Test for Undiagnosed Hyperglycemia in Ethnic Minority Groups Diabetes Care 2004;27(1):116-122.

(51) Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabetic Medicine 2010 Aug;27(8):887-895.

(52) Gray L, Khunti K, Edwardson C, Goldby S, Henson J, Morris D, et al. Implementation of the automated Leicester Practice Risk Score in two diabetes prevention trials provides a high yield of people with abnormal glucose tolerance. Diabetologia 2012;55(12):3238-3244.

(53) van den Donk M, Sandbaek A, Borch-Johnsen K, et al. Screening for type 2 diabetes. Lessons from the ADDITION-Europe study. Diabetic Medicine 2011;28:1416-1424.

(54) Abbasi A, Peelen L, Corpeleijn E, van der Schouw Y, Stolk R, Spijkerman A. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. BMJ 2012(345).

(55) Brown N, Critchley J, Bogowicz P, Mayigee M, Unwin N. Risk scores based on self-reported or available clinical data to detect undiagnosed Type 2 Diabetes: A systematic review. Diabetes Research and Clinical Practice 2012;98(3):369-385.

(56) Buijsse B, Simmons R, Griffin S, Schulze M. Risk Assessment Tools for Identifying Individuals at Risk of Developing Type 2 Diabetes. Epidemiologic Reviews 2011;33(1):46--62.

(57) Collins G, Mallett S, Omar O, Yu L. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine 2011;9(1):103.

(58) Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011;343.

(59) Thoopputra T, Newby D, Schneider J, Li S. Survey of diabetes risk assessment tools: concepts, structure and performance. Diabetes/Metabolism Research and Reviews 2012;28(6):485-498.

(60) Breiman L, Friedman R, Olshen C, Stone J. Classification and Regression Trees. New York: Chapman & Hall/CRC; 1984.

(61) Dietterich T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning 2000;40(2):139-157.

(62) Heikes KE, Eddy DM, Arondekar B, Schlessinger L. Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes Care 2008 May;31(5):1040-1045.

(63) Wilson P, Meigs J, Sullivan L, Fox C, Nathan D, D'Agostino R. Prediction of Incident Diabetes Mellitus in Middle-aged Adults. JAMA Internal Medicine 2007;167(10):1068-1074.

(64) Royston P, Altman D, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine 2006;25(1):127-141.

(65) Janssen K, Siccama I, Vergouwe Y, Koffijberg H, Debray T, Keijzer M, et al. Development and validation of clinical prediction models: marginal differences between logistic regression, penalized maximum likelihood estimation, and genetic programming. Journal of Clinical Chemistry 2012;65(4):404-412.

(66) Flom P, Cassell D. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use 2007; Available at: http://www.nesug.org/proceedings/nesug07/sa/sa07.pdf. Accessed 08/06, 2013.

(67) Harrell F. Regression Modelling Strategies. 2nd Ed.: New York:Springer; 2002.

(68) Efron B, Hastie T, Tibshirani R. Least angle regression. Annals of Statistics 2004;32:407-499.

(69) Weisberg S. Discussion of “Least Angle Regression” by Efron et al. The Annals of Statistics 2004;32(2):490-494.

(70) Harrell FJ, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 1996;15:361-387.

(71) Sun G, Shook T, Kay G. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of clinical epidemiology 1996;49:907-916.

(72) Collins G, de Groot J, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology 2014;14(40).

(73) Hyam J, Welch C, Harrison D, Menon D. Case mix, outcomes and comparison of risk prediction models for admissions to adult, general and specialist critical care units for head injury: a secondary analysis of the ICNARC Case Mix Programme Database. Critical Care 2006;10:S2.

(74) Kramer A, Zimmerman J. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Critical Care Medicine 2007;35(9):2052-2056.

(75) Altman D, Royston P. What do we mean by validating a prognostic model?. Statistics in Medicine 2000;19:453-473.

(76) Royston P, Altman D. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Notes 2013;13(33).

(77) Witte DR, Shipley MJ, Marmot MG, Brunner EJ. Performance of existing risk scores in screening for undiagnosed diabetes: an external validation study. Diabetic Medicine. 2010;27(1):46-53.

(78) Vergouwe Y, Moons K, Steyerberg E. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. American Journal of Epidemiology 2010;172(8):971-980.

(79) McNeely M, Boyko EJ, Leonetti D, Kahn S, Fujimoto W. Comparison of a Clinical Model, the Oral Glucose Tolerance Test, and Fasting Glucose for Prediction of Type 2 Diabetes Risk in Japanese Americans. Diabetes Care 2003;26(3):758-763.

(80) Herman W. Predicting risk for diabetes: choosing (or building) the right model. Annals of Internal Medicine 2009;150(11):812-814.

(81) Absetz P, Oldenburg B, Hankonen N, Valve R, Heinonen H, Nissinen A. Type 2 diabetes prevention in the real world: three-year results of the GOAL lifestyle implementation trial. Diabetes Care 2009;32:1418-1420.

(82) Absetz P, Oldenburg B, Hankonen N, Valve R, Heinonen H, Nissinen A. Type 2 diabetes prevention in the real world: three-year results of the GOAL lifestyle implementation trial. Diabetes Care 2009;32:1418-1420.

(83) Kulzer B, Hermanns N, Gorges D, Schwarz P, Haak T. Prevention of diabetes self-management program (PREDIAS): effects on weight, metabolic risk factors, and behavioral outcomes. Diabetes Care 2009;32:1143-1146.

(84) Laatikainen T, Dunbar J, Chapman A, et. al. Prevention of type 2 diabetes by lifestyle intervention in an Australian primary health care setting: Greater Green Triangle (GGT) Diabetes Prevention Project. BMC Public Health 2007;7:249.

(85) Lindström J, Absetz P, Hemio k, Peltomaki P, Peltomen M. Reducing the risk of type 2 diabetes with nutrition and physical activity—efficacy and implementation of lifestyle interventions in Finland. Public Health Nutrition 2010;13(6A):993-999.

(86) Moons K, Kengne A, Woodward M, Royston P, Vergouwe Y, Altman D, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683-690.

(87) Moons K, Kengne A, Grobbee D, Royston P, Vergouwe Y, Altman D, et al. Risk prediction models: II. External validation, model updating, and impact assessment Heart 2012;98:691-698.

(88) Webb DR, Khunti K, Srinivasan B, Gray L, Taub N, Campbell S, et al. Rationale and design of the ADDITION-Leicester study, a systematic screening programme and Randomised Controlled Trial of multi-factorial cardiovascular risk intervention in people with Type 2 Diabetes Mellitus detected by screening. Trials 2010;11(16).

(89) Webb DR, Gray L, Khunti K, Srinivasan B, Taub N, Campbell S, et al. Screening for diabetes using an oral glucose tolerance test within a western multi-ethnic population identifies modifiable cardiovascular risk: the ADDITION-Leicester study. Diabetologia 2011;54:2237-2246.

(90) Alberti K, Zimmet P. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. 1998.

(91) Batty D, Blake M, Bridges S, Crawford R, Demakakos P, de Oliveira C, et al. The dynamics of ageing: Evidence from the English Longitudinal Study of Ageing 2002-12 (Wave 6). Peterborough: Printondemand-worldwide.com; 2014.

(92) Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: the English longitudinal study of ageing. International Journal of Epidemiology 2013;42(6):1640-1648.

(93) Banks J, Nazroo J, Steptoe A. The dynamics of ageing: evidence from the English longitudinal study of ageing 2002–12 (wave 6). London: The Institute for Fiscal Studies; 2014.

(94) UK Data Service. Discover Data. 2016; Available at: https://www.ukdataservice.ac.uk/get-data/.

(95) Parikh R, Mathai A, Parikh S, Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian Journal of Opthalmology 2008;56(1):45-50.

(96) Altman D, Bland J. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ 1994;308:1552.

(97) Altman D, Bland J. Diagnostic tests 2: Predictive values. BMJ 1994;309(6947):102.

(98) Altman D, Bland J. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots. BMJ 1994;309:188.

(99) Barber S, Davies M, Khunti K, Gray L. Risk assessment tools for detecting those with pre-diabetes: A systematic review. Diabetes Research and Clinical Practice 2014;105(1):1-13.

(100) Brier G. Verification of forecasts expressed in terms of probability. Monthly weather review 1950;78(1):1-3.

(101) Blattenberger G, Lad F. Separating the Brier score into calibration and refinement components: A graphical exposition. The American statistician 1985;39(1):26-32.

(102) Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 2010;21(1):128-138.

(103) Pencina M, D'Agostino RS, D'Agostino RJ, Vasan R. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in Medicine 2008;27(2):157-172.

(104) Khunti K, Gillies C, Taub N, Mostafa S, Hiles S, Abrams k, et al. A comparison of cost per case detected of screening strategies for Type 2 diabetes and impaired glucose regulation: Modelling study. Diabetes Research and Clinical Practice 2012;97(3):505-513.

(105) Handlos LN, Witte DR, Almdal TP, Nielsen LB, Badawi SE, Sheikh ARA, et al. Risk scores for diabetes and impaired glycaemia in the Middle East and North Africa. Diabetic Medicine 2012;30(4):443-451.

(106) Yu W, Liu T, Valdez R, Gwinn M, Khoury M. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 2010;10(16).

(107) Xin Z, Yuan J, Hua L, Ma YH, Zhao L, Lu Y, et al. A simple tool detected diabetes and prediabetes in rural Chinese. Journal of Clinical Epidemiology 2010;63(9):1030-1035.

(108) Barriga KJ, Hamman RF, Hoag S, Marshall JA, Shetterly SM. Population screening for glucose intolerant subjects using decision tree analyses. Diabetes Research & Clinical Practice 1996;34(S17-29).

(109) DuBose KD, Cummings DM, Imai S, Lazorick S, Collier DN. Development and validation of a tool for assessing glucose impairment in adolescents. Preventing Chronic Disease 2012;9:E104.

(110) Hische M, Luis-Dominguez O, Pfeiffer AF, Schwarz PE, Selbig J, Spranger J. Decision trees as a simple-to-use and reliable tool to identify individuals with impaired glucose metabolism or type 2 diabetes mellitus. European Journal of Endocrinology 2010;163(4):565-571.

(111) Koopman RJ, Mainous AG,3rd, Everett CJ, Carter RE. Tool to assess likelihood of fasting glucose impairment (TAG-IT). Annals of Family Medicine 2008;6(6):555-561.

(112) Nelson KM, Boyko EJ, Third National Health and Nutrition Examination,Survey. Predicting impaired glucose tolerance using common clinical information: data from the Third National Health and Nutrition Examination Survey. Diabetes Care 2003;26(7):2058-2062.

(113) Robinson CA, Agarwal G, Nerenberg K. Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population. Chronic Diseases and Injuries in Canada 2011 Dec;32(1):19-31.

(114) Quinlan JR. C4.5: Programs for Machine Learning. 1st Ed.: Morgan Kaufmann, Los Altos; 1993.

(115) Gray L, Khunti K, Wilmot E, Yates T, Davies M. External validation of two diabetes risk scores in a young UK South Asian population. Diabetes Research & Clinical Practice 2014;104(3):451-458.

(116) Heianza Y, Arase Y, Saito K, Hsieh S, Tsuji H, Kodama S, et al. Development of a Screening Score for Undiagnosed Diabetes and Its Application in Estimating Absolute Risk of Future Type 2 Diabetes in Japan: Toranomon Hospital Health Management Center Study 10 (TOPICS 10). The Journal of Clinical Endocrinology & Metabolism 2013;98(3).

(117) Piette J, Milton E, Aiello A, Mendoza-Avelares M, Herman W. Comparison of three methods for diabetes screening in a rural clinic in Honduras. Pan American Journal of Public Health 2010;28(1):49-57.

(118) Phillips L, Ziemer D, Kolm P, Weintraub W, Vaccarino V, Rhee M, et al. Glucose challenge test screening for prediabetes and undiagnosed diabetes. Diabetologia 2009;52(9):1798-1807.

(119) Lee Y, Bang H, Kim H, Kim H, Park S, Kim D. A Simple Screening Score for Diabetes for the Korean Population. Diabetes Care 2012;35(8):1723-1730.

(120) Rorie J, Smith A, Evans T, Horsburgh R, Brooks D, Goodman R, et al. Using Resident Health Advocates to Improve Public Health Screening and Follow-Up Among Public Housing Residents, Boston, 2007-2008. Preventing Chronic Disease 2011;8(1):A15.

(121) Jones A, Knight B, Baker G, Hattersley A. Practical implications of choice of test in National Institute for Health and Clinical Excellence (NICE) guidance for the prevention of Type 2 diabetes. Diabetic Medicine 2012;30(1):126-127.

(122) Miller A. Subset selection in regression. 2nd Ed.: Chapman & Hall, London; 2002.

(123) Steyerberg E, Bleeker S, Moll H, Grobbee D, Moons K. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. Journal of Clinical Epidemiology 2003;56(5):441-447.

(124) Caruana R. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference of Machine learning; 2006.

(125) Cooper G, Aliferis C, Ambrosino R, Aronis J, Buchanan B, Caruana R, et al. An evaluation of machine learning methods for predicting pneumonia mortality. Artificial Intelligence in Medicine 1997;9.

(126) Lehmann C, Koenig T, Jelic V, Prichep L, John R, Wahlund L, et al. Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). 2007;161(2):342-350.

(127) Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Research Notes 2011;4(299).

(128) Tapak L, Hossein M, Hamidi O, Poorolajal J. Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran. Healthcare Informatics Research 2013;19(3):177-185.

(129) Bouwmeester W, Zuithoff N, Mallett S, Geerlings M, Vergouwe Y, Steyerberg E, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. PLOS Medicine 2012.

(130) Little R, Rubin D. Statistical analysis with missing data. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2002.

(131) Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.

(132) Harrell F. Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. 2nd Ed.: New York: Springer; 2001.

(133) Donders A, van der Heijden G, Stijnen T, Moons K. Review: a gentle introduction to imputation of missing values. Journal of Clinical Epidemiology 2006;59:1087-1091.

(134) Greenland S, Finkle W. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology 1995;142:1255-1264.

(135) Gorelick M. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Journal of Clinical Epidemiology 2006;59:1115-1123.

(136) Sterne J, White I, Carlin J, Spratt M, Royston P, Kenward M, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;29(338).

(137) Janssen K, Donders A, Harrell F, Vergouwe Y, Chen Q, Grobbee Q, et al. Missing covariate data in medical research: To impute is better than to ignore. Journal of Clinical Epidemiology 2010;63:721-727.

(138) Rubin D. Multiple Imputation for Nonresponse in Surveys. The Proceedings of the Survey Research Methods Section of the American Statistical Association 1978:20-34.

(139) Collins L, Schafer J, et al. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 2001;6(4):330-351.

(140) Moons K, Donders R, Stijnen T, Harrell FJ. Using the outcome for imputation of missing predictor values was preferred. Journal of clinical epidemiology 2006;59(10):1092-1101.

(141) White I, Royston P, Wood A. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 2011;30(4):377-399.

(142) Van Buuren S, Boshuizen H, Knook D. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 1999;18(6):681-694.

(143) Raghunathan T, Lepkowski J, Van Hoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001;7(4):445-464.

(144) Lee K, Carlin J. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology 2010;171(5):624-632.

(145) Graham J, Olchowski A, Gilreath T. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science 2007;8(3):206-213.

(146) StataCorp. *Stata Statistical Software: Release 13*. 2013.

(147) Royston P. Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. Stata Journal 2009;9:466-477.

(148) Gould W, Pitblado J, Poi P. Maximum Likelihood Estimation with Stata. 4th ed.: Stata Press, 2010.

(149) Hauck W, Donner A. Wald's Test as Applied to Hypotheses in Logit Analysis. Journal of the American Statistical Association 1977;72:851-853.

(150) Derksen S, Keselman H. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables**.** British Journal of Mathematical and Statistical Psychology 1992;45(2): 265–282.

(151) Tan P, Steinbach M, Kumar V. Classification: basic concepts, decision trees and model evaluation. Introduction to data mining. 1st ed. Boston, MA: Pearson Addison Wesley; 2006. p. 145-205.

(152) Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. Available at: https://cran.r-project.org/web/packages/rpart/rpart.pdf Accessed 04/07, 2015**.**

(153) Efron B. Estimating the error rate of a prediction rule improvement on cross-validation. Journal of the American Statistical Association 1983;78(382):316-330.

(154) Archer K. rpart ordinal: An R Package for Deriving a Classification Tree for Predicting an Ordinal Response. Journal of Statistical Software 2010;34(7).

(155) Friedman J. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 2001;29(5):1189-1232.

(156) Elith J, Leathwick J, Hastie T. A working guide to boosted regression trees. Journal of Animal Ecology 2008;77(4):802-813.

(157) Hastie T, Tibshirani R, Friedman J. 10. Boosting and Additive Trees. The elements of statistical learning. 2nd ed.: New York: Springer; 2009. p. 337-384.

(158) Friedman J. Stochastic gradient boosting Nonlinear methods and Data Mining 2002;38(4):367-378.

(159) Ridgeway G, et al. gbm: Generalized Boosted Regression Models. Available at: https://cran.r-project.org/web/packages/gbm/gbm.pdf Accessed 04/07, 2015.

(160) Breiman L. Bagging Predictors. Machine Learning 1996;24:123-140.

(161) Peters A, Hotthorn T. ipred: Improved Predictors. Available at: https://cran.r-project.org/web/packages/ipred/ipred.pdf Accessed 05/06, 2014

(162) Aslam JA, Popa RA Rivest RL. On Estimating the Size and Confidence of a Statistical Audit. Proceedings of the Electronic Voting Technology Workshop (EVT '07); 2007.

(163) Breiman L. Random Forests. Machine Learning 2001;45(1):5-32.

(164) Ho Tk. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998;20(8):832-844.

(165) Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2(3):18-22.

(166) Boser BE, et al. A Training Algorithm for Optimal Margin Classifiers. The Fifth Annual Workshop on Computational Learning Theory; 1992.

(167) Cover T. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. Electronic Computers 1965;3:326-334.

(168) Cortes C, Vapnik V. Support-Vector Networks. Machine Learning 1995;20:273-297.

(169) Meyer D, et al. Misc Functions of the Department of Statistics (e1071). 2012. Available at: https://cran.r-project.org/web/packages/e1071/e1071.pdf Accessed 12/09, 2015.

(170) Platt J. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. Advances in Kernel Methods: Support Vector Learning: MIT press; 1999. p. 185-210.

(171) Webb DR, Khunti K, Srinivasan B, Gray L, Jarvis J, Hiles S, et al. ADDITION Leicester: prevalence of impaired glucose regulation (IGR) and screen detected Type 2 diabetes (T2D) in a mixed ethnic UK population. Diabetic Medicine 2009;26 (S1)(3).

(172) Logue J, Walker J, Colhoun H, Leese G, Lindsay M, JA., et al. Do men develop type 2 diabetes at lower body mass indices than women? Diabetologia 2011;54(12):3003-3006.

(173) Harrell F, Lee K, Matchart D, Reichert T. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep 1985;69:1071-1077.

(174) Moons K, Royston P, Vergouwe Y, Grobbee D, Altman D. Prognosis and prognostic research: what, why, and how? BMJ 2009;338(b375).

(175) Peduzzi P, Concato J, Kemper E, Holford T, Feinstein A. A simulation study on the number of events per variable in logistic regression analysis. Journal of clinical epidemiology 1996;49:1373-1379.

(176) Vittinghoff E, McCulloch C. Relaxing the rule of ten events per variable in logistic and Cox regression. American Journal of Epidemiology 2007;165:710-718.

(177) Steyerberg E, Schemper M, Harrell F. Logistic regression modelling and the number of events per variable: selection bias dominates. Journal of clinical epidemiology 2011;64:1463-1469.

(178) Steyerberg E, Eijkemans M, Habbema J. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. Journal of clinical epidemiology 1999;52:935-942.

(179) Harrell F. rms: Regression Modeling Strategies Available at: https://cran.r-project.org/web/packages/rms/index.html Accessed 02/02, 2017.

(180) Pedersen J, Gerds T, Bjorner J, Christensen K. Prediction of future labour market outcome in a cohort of long-term sick-listed Danes. BMC Public Health 2014;14:494.

(181) Ogundimu E, Altman D, Collins G. Adequate sample size for developing prediction models is not simply related to events per variable. Journal of clinical epidemiology 2016;76:175-182.

(182) Smith J, Anderson P. Conditional independence and chain event graphs. Artifical Intelligence 2008;172:42-68.

(183) Shafer G, Gillett P, Scherl R. A new understanding of subjective probability and its generalization to lower and upper prevision. International Journal of Reason 2003;33(1):1-49.

(184) Freeman G, Smith J. Bayesian MAP model selection of chain event graphs. Journal of Multivariate Analysis 2011;102(7):1152-1156.

(185) Thwaites P. Chain Event Graph MAP model selection. KEOD 2009: proceedings of the international conference on knowledge engineering and ontology development; 2009.

(186) Barclay L, Hutton J, Smith J. Embellishing a Bayesian Network using a Chain Event Graph. CRiSM Paper 2012:10/08.

(187) Collins G, Ogundimu E, Cook J, Le Manach Y, Altman D. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Statistics in Medicine 2016.

(188) Collazo R, Smith J. A New Family of Non-Local Priors for Chain Event Graph Model Selection. Bayesian Analysis 2016.

(189) Lagakos S. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Statistics in Medicine 1988;7:257-274.

(190) Pencina K, Pencina M, D'Agostino RS. What to expect from net reclassification improvement with three categories. Statistics in Medicine 2014;33(28):4975-4987.

(191) Kerr K, Wang Z, Janes H, McClellend R, Psaty B, Pepe M. Net Reclassification Indices for Evaluating Risk-Prediction Instruments: A Critical Review. Epidemiology 2014;25(1):114-121.

(192) Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20(1):37-46.

(193) Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

(194) Diabetes UK. Type 2 Diabetes: Know your risk. 2016; Available at:http://bit.ly/23KrbzE. Accessed 02/09, 2016.

(195) Office for National Statistics. National population projections 2004-based. 2006.

(196) Pattison J, McPHerson K, Blakemore C, Haberman S, et al. Life expectancy: Past and future variations by gender in England and Wales. Longevity Science Advisory Panel 2012(2).

(197) Gray B, Bracken R, Turner D, Morgan K, Thomas M, Williams S, et al. Different type 2 diabetes risk assessments predict dissimilar numbers at 'high risk': a retrospective analysis of diabetes risk-assessment tools. British Journal of General Practice 2015;65(641):852-860.

(198) Collins G, Ogundimu E, Altman D. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in Medicine 2016;35(2):214-226.

(199) Vergouwe Y, Steyerberg E, Eijkemans M, Habbema J. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of Clinical Epidemiology 2005;58:475-483.

(200) Schwarz P, Muller G. How to screen for diabetes risk in multi-ethnic populations: does one method fit all? European Diabetes Nursing 2013;10(2):63-68.

(201) Lee C, Colagiuri S. Risk scores for diabetes prediction: The International Diabetes Federation PREDICT-2 project. Diabetes Research and Clinical Practice 2013;100(2):285-286.

(202) Ahmed I, Debray T, Moons K, Riley R. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Medical Research Methodology 2014;14(3).

(203) Riley R, Lambert P, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340.

(204) International Diabetes Federation. Risk prediction tools (PREDICT - 2). 2015; Available at: http://www.idf.org/risk-prediction-tools-predict-2. Accessed 07/24, 2016.

(205) Akbarov A, Williams R, Brown B, Mamas M, Peek N, Buchan I, et al. A Two-stage Dynamic Model to Enable Updating of Clinical Risk Prediction from Longitudinal Health Record Data: Illustrated with Kidney Function Studies in Health and Informatics 2015;216:696-700.

(206) Crowther M, Abrams K, Lambert P. Joint modelling of longitudinal and survival data. Stata Journal 2013;13:165-184.