Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank

1.1 Author names

Jacob Armstrong[1] [0000-0003-4349-4453], Justine K. Rudkin[1] [0000-0002-3044-0459], Naomi Allen[1,2] [0000-0003-1938-5038], Derrick W. Crook[3] [0000-0002-0590-2850], Daniel J. Wilson[1] [0000-0002-0940-3311], David H. Wyllie[3,4]* [0000-0002-9155-6228] and Anne-Marie O'Connell[4]**

** Joint corresponding authors*

1.2 Affiliation

1. *Big Data Institute, Nuffield Department of Population Health, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford, OX3 7LF, United Kingdom*
2. *UK Biobank, 1-2 Spectrum Way, Adswood, Stockport, SK3 0SA, United Kingdom*
3. *Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom*
4. *Public Health England, Field Service, Addenbrookes' Hospital, Cambridge CB2 0QQ, United Kingdom*

1.3 Corresponding author

daniel.wilson@bdi.ox.ac.uk, david.wyllie@phe.gov.uk, anne-marie.oconnell@phe.gov.uk

1.4 Keyword

UK Biobank, Public Health England, Microbiology, Infection, Database linkage, Covid-19, SARS-CoV-2, Bugbank

1.5 Repositories:

Not applicable

2. Abstract

UK Biobank is an international health resource enabling research into the genetic and lifestyle determinants of common diseases of middle and older age. It comprises 500,000 participants.

Public Health England's Second Generation Surveillance System is a centralized microbiology database covering English clinical diagnostics laboratories that provides national surveillance of legally notifiable infections, bacterial isolations and antimicrobial resistance. We previously developed secure, anonymized, individual-level linkage of these systems. In this study, we implemented rapid dynamic linkage, which allows us to provide a regular feed of new COVID-19 (SARS-CoV-2) test results to UK Biobank to facilitate rapid and urgent research into the epidemiological and human genetic risk factors for severe infection in the cohort. Here, we have characterized the first 678 cases of COVID-19 in UKB participants, of whom 552 met our working definition of severe COVID-19 as inpatients hospitalized on or after 16 March 2020. We found that the incidence of severe COVID-19 among UKB cases was 27.1% lower than the general population in England, although this difference varied significantly by age and sex. The total number of UKB cases comprised could be estimated as 0.9% of the publicly announced number of cases in England. We considered how increasing case numbers will affect the power of genome-wide association studies. This new dynamic linkage system has further potential to facilitate the investigation of other infections and the prospective collection of microbiological cultures to create a microbiological biobank ('bugbank') for studying the interaction of environment, human and microbial genetics on infection in the UK Biobank cohort.

3. Impact statement

Infections are a major source of human disease around the world, both during outbreaks such as the ongoing COVID-19 pandemic, and in ordinary times. Scientific research provides the foundation of new knowledge about the risks and consequences of infection. This research can contribute to delivering new drugs and vaccines and to public health policy. In this article we report our contribution to facilitating research into the risk factors for severe COVID-19 and other infectious diseases by integrating information between two valuable resources: the UK Biobank and a Public Health England national microbiology database. UK Biobank involves citizens who have provided consent for their de-identified data to be accessed by approved researchers worldwide to perform health research that is in the public interest. Beginning in 2006, the study recruited men and women aged 40-69 years across the UK and collected a vast array of lifestyle data, physical measures and biological samples (for genomic and other assays to be performed). These data, together with long-term linkage to their electronic medical records, provide an unprecedented resource to understand the epidemiology of diseases of middle and older age. In this article we report a new computerized system that provides daily linkage of participants with their microbiological test results, with the aim of providing data about COVID-19 and other infections in the UK Biobank cohort.

4. Data summary

The code written for database linkage in this study is internal to Public Health England systems, and will not be released publicly. The data provided by the PHE system will be incorporated into the UK Biobank database and released through the usual channels. To access UK Biobank data, researchers must register and submit a research application (https://www.ukbiobank.ac.uk/register-apply). Registration is open to all bona fide researchers for all types of health-related research that is in the public interest. The registration and

application process ensures researchers and projects meet UK Biobank's obligations to its participants and funders.

5. Introduction

As of 14 April 2020, the coronavirus SARS-CoV-2 that causes the severe acute respiratory syndrome COVID-19 was reported to have infected over 2 million people and killed over 120,000 people around the world (Zhu et al 2020; Dong, Du and Gardner, 2020). Better understanding of this novel pathogen is urgently needed to help guide the improvement of treatment and prevention. Large cohort studies such as the UK Biobank, which have gathered detailed epidemiological, medical and genetic records of hundreds of thousands of people, offer the opportunity to uncover risk factors for COVID-19, including the molecular genetic pathways underlying severe disease.

UK Biobank (UKB) is a longitudinal prospective cohort study that aims to investigate the causes, treatment and prevention of many common diseases of middle and older age (Sudlow et al 2015). The cohort is a particularly appropriate focus for the study of COVID-19 because incidence of this severe disease increases with age (Chen et al 2020, Chinese Preventive Medicine Association 2020, Huang et al 2020, Li et al 2020, Novel Coronavirus Pneumonia Emergency Response Epidemiology Team 2020, Wang et al 2020, Yang et al 2020, Zhu et al 2020). The UKB cohort comprises around 500,000 men and women from the United Kingdom who were aged 40-69 years when they were recruited in 2006-2010; 427,000 individuals were still being followed up at the end of 2019. Participants attended assessment centres, provided detailed information on lifestyle and medical history, underwent a range of physical measures and provided biological samples for future assays. They also provided consent for UKB to follow their health over the longer term by linking to their health-related records. Research scientists around the world can register and apply for access to UKB data, allowing them to study lifestyle, environmental and human genetic risk factors for disease (www.ukbiobank.ac.uk).

Studies of infection within UKB mainly rely on identifying infection events among participants from electronic medical records. To date, this predominantly comes from hospital inpatient admissions, including the hospital episode statistics (HES), which contain diagnoses assigned by professional coding teams post-discharge based on medical records. Acute diagnoses, and those of underlying conditions, are codified using the ICD-10 (international classification of disease) system. Emergency ICD-10 codes for COVID-19 have been assigned (U07.1 COVID-19, virus identified; U07.2 COVID-19, virus not identified; https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf). However, there are limitations to these data for studying infection, as coding occurs in local National Health Service (NHS) Hospital Trusts, with subsequent central collation by the NHS and periodic (currently monthly) incorporation of summaries into UK Biobank. Other limitations of HES for studying infection include incomplete or insensitive microbiological testing, and difficulty in syndromic diagnosis, especially in the elderly (e.g. Yoshikawa 2000), where infection can exacerbate pre-existing conditions, so that not all causes of infection-related hospitalization are necessarily recorded as such. Moreover, infection diagnosis, and its coding, is often imprecise: for example ICD-10 permits broad non-specific categories to be recorded such as A41.9 "Septicaemia, unspecified", of which there were 2660 cases among UKB participants by 2017.

The Public Health England (PHE) microbiology database SGSS (Second Generation Surveillance System) offers advantages over HES data for the ascertainment of infection in UKB participants because it provides more granular and highly specific diagnosis of microbiological confirmed infection, including both COVID-19 and infections caused by micro-organisms with antimicrobial resistance. Of note, it only allows identification of microbiologically confirmed disease; in the case of COVID-19, other databases with a more clinical focus also exist, such as the PHE CHESS and Intensive Care National Audit & Research Centre (ICNARC) databases containing individual patient data on critically ill patients in intensive care units. The SGSS database provides good coverage of UKB participants, as most (89%) of them resided in England at recruitment. For these reasons, we previously developed secure, anonymized, individual-level linkage between SGSS and UKB with a view to providing data feeds periodically, e.g. annually, as with other data sources like cancer registries (Hilton et al. 2020).

The NHS and PHE have put in place microbiological testing for SARS-CoV-2. As of 13 April 2020, 89 laboratories were reporting positive SARS-CoV-2 nucleic acid detection results to SGSS, of which 76 were reporting negative results as well; over 186,000 positive tests had been reported. More laboratories are coming on line, with the aim that all laboratories in England report both positive and negative results. As well as playing a critical role in patient diagnosis, these data are important to enable research into the epidemiological and genetic determinants of severe COVID-19. In this paper, we report the development of a dynamic linkage system that identifies new records in SGSS from UKB participants on a daily basis, and feeds those results back to UKB weekly.

We originally developed this system as a pilot study to determine the feasibility of prospective microbiological culture collection from UKB participants to create a microbiological biobank ('bugbank') for joint studies of epidemiological, human genetic and pathogen genomic risk factors for infection. In light of the COVID-19 pandemic, we have repurposed the system to provide near-to-real-time data on SARS-CoV-2 positive and negative test results for UKB participants. Here we characterize the first 678 identified cases in the cohort and compare their demographic characteristics to the rest of the UK Biobank cohort and to other cases in England.

6. Dynamic data linkage

We established a dedicated server at PHE Colindale to manage dynamic linkage between UKB and SGSS. All NHS microbiological laboratories in England provide data to SGSS each working day. SGSS consumes two data feeds, performing quality control checks and applying mappings between terms used by individual laboratories to produce a standardized dataset. The antimicrobial resistance (AMR) feed contains data from all microbiological cultures on which AMR testing was performed. The communicable disease report (CDR) feed contains mandatory reporting of a narrow range of pathogens of particular public health importance, including SARS-CoV-2. Our algorithms link to both the AMR feed for the prospective micro-organism retrieval pilot study and the CDR feed for the COVID-19 rapid response project.

There are specific challenges which arise when frequently linking data between the large SGSS and UKB participant databases. These challenges pertain to the computational demands of

dealing with high-volume, high-frequency queries. Building on our previous static linkage approach (Hilton et al 2020), we developed a speed optimized algorithm (e.g. Shin and Sanders, 2006) to implement incremental daily linkage of the circa 200,000 records fed into SGSS each day and to identify those belonging to UKB participants.

The key steps in our system, summarized in Figure 1, are:

1. An agent runs persistently on a server at PHE Colindale hosting SGSS receiving daily updates from NHS/PHE laboratories across the country.
2. Periodically, the agent updates a database held at PHE Colindale with any new records from SGSS that it matches with UK Biobank participants.

   The record matching procedure uses computerized pseudonymization (OpenPseudonymiser) to maintain privacy and prevent inadvertent disclosure of patient identifiers, as previously described (Hilton et al 2020).
3. Periodically, an extract of the data is transferred to UK Biobank for ingestion into their systems.

To enable prospective culture collection feasibility study, the further steps undertaken are:

4. The agent sends an email alert to the key person, e.g. a Biomedical Scientist, at the NHS/PHE laboratory to alert them that new samples have arrived (Figure 2).
5. The key person accesses the details of the microbiological cultures necessary for retrieval, retrieves the identified sample and, if appropriate, makes a stock of the microbial growth for freezer storage. Each sample is assigned a unique sample identifier and storage location which is logged into the secure system. The key person logs any samples that could not be located and other non-personal information relating to the sample relevant to the pilot goals, such as noting physical damage, lack of growth or low growth.

To test the functionality of the system, we implemented an automated bi-daily SGSS-UKB cohort linkage, with automated daily email alerts of any new records. The prospective pilot study was commenced, with email alerts describing which samples to target sent to the relevant laboratory through the PHE secure network (Figure 2). We have tested this system by collecting bacteriology samples in the John Radcliffe Hospital, Oxford, prior to proposed England-wide deployment, although it could also be applied to COVID-19 samples. Figure 3 summarizes the UKB participants' infection events reported by the microbiology lab of the John Radcliffe Hospital, Oxford in September 2019, which is among the 10 English labs with the most frequent UKB infections. The outcomes of the pilot study, including retrieval rates, will be presented in future work, but initial results indicate that the dynamic record system does allow us to retrieve samples in a timely manner before specimens are discarded, as is routine in microbiology laboratories. A blanket rule requiring COVID-19 positive material to be retained for three days would likely enable a very high retrieval rate of positive samples.

7. Characterization of first detected cases of COVID-19 in UKB

7.1 Working definition for identifying severe COVID-19 cases in UKB

A key question for the international research community is what factors predispose individuals to severe COVID-19? We considered whether these individuals could be identified from the data available. Although SGSS does not contain clinical illness severity (this will come from linking UKB to electronic medical records), SGSS does contain the origin of samples. This is relevant because, from 16 March 2020, the UK entered a suppression phase aimed at delaying the outbreak, during which COVID-19 testing was largely restricted to inpatients, and hospitalisation was restricted to those requiring medical support. Indeed, even access to accident and emergency (A&E) departments for patients with suspected COVID-19 requires assessment by a telephone service (111), which only refers severe cases to hospital. In contrast, during the preceding containment phase, referral to hospital was practised even for those with very mild disease for infection control reasons.

Therefore, for individuals sampled from 16 March 2020, we propose that testing positive for SARS-CoV-2 as a hospital inpatient is an appropriate surrogate of severe disease for initial analyses. This definition is not necessarily sensitive, as individuals tested in the community and subsequently admitted are not included in our definition unless they are re-tested in hospital and found positive.

From SGSS, we define a *hospital inpatient* as any person with at least one positive SARS-CoV-2 test having one or more of the following properties:

- The Requesting Organisation Type associated with the test was either “Hospital Inpatient” or “Hospital A&E”.
- The record possessed an “Acute Trust” flag, meaning that the test came from a hospital delivering emergency care, or
- The record possessed a “Hospital Acquired Infection” flag.

Excluding those who are being tested because they are a ‘Healthcare Worker' (i.e. the associated Requesting Organisation Type is “Healthcare Worker Testing”).

Manual curation of a sample of records identified by this method, and inspection of free-text information about ward or other sampling locations, indicated that this definition provided a specific means of identifying inpatient samples.

All other patients with positive SARS-CoV-2 test results in SGSS are still included in the data feed to UKB; we consider them to be *non-inpatients*. Some of these *non-inpatient* records may be from individuals subsequently admitted to hospital, and so the *non-inpatient* designation does not necessarily reflect mild disease.

Of note, future integration into UK Biobank of HES and ICU data which record severity and augmented care periods information (e.g. intensive and high dependency wards stays) is planned in the future. Therefore, more refined classifications may be provided in later releases of the UKB data.

7.2 Demographic characteristics of UKB participants with severe COVID-19

As of 14 April 2020, 552 UKB participants reported a positive SARS-CoV-2 test while hospital inpatients in England on or after 16 March 2020. A further 126 participants were classified as non-inpatients (102 cases). We excluded from further analysis 24 cases because they had their first positive SARS-CoV-2 test before 16th March 2020.

The total number of PCR-confirmed COVID-19 cases between 16th March and 10th April (allowing a reporting lag of 5 days) in England, reported by the UK Government, was 53,455 (data from https://coronavirus.data.gov.uk). The total number of UKB participants with SARS-CoV-2 positive tests in SGSS over the same period was 590, of whom we classed 507 as inpatients. Thus, UKB participants meeting our operational definition of severe (i.e. hospitalised) COVID-19 disease currently comprise 0.9% of the total number of COVID-19 cases reported in public UK Government data for England.

The number of new cases of COVID-19 inpatients recorded in SGSS has increased rapidly since early 2020, with close correspondence between the growth in cases among UKB participants and England as a whole (Figure 4). The total number of inpatients with SARS-CoV-2 positive tests between 16 March and 10 April in England, recorded by SGSS, was 52,049. Thus, UKB participants made up 1.0% of all COVID-19 inpatients in SGSS. The change in this proportion over time was not significantly different to zero, although there was substantial statistical uncertainty (95% CI -1.9-5.8% per day, Poisson regression). Thus, the outbreak dynamics appear similar between UKB participants and the general population of England.

We estimated per capita incidence of our operational definition of severe COVID-19 for the period 16th March – 14th April using SGSS data, not restricted to UKB participants (Figure 5A). In keeping with other reports (see Introduction), males were generally at elevated risk of developing severe COVID-19 across England (odds ratio 1.29, 95% CI 1.27-1.32; Cochran-Mantel-Haenszel test stratified by age group), relative to females. We compared UKB cases to all cases in England using SGSS data to test for systematic differences in incidence by age and sex. Because of the recruitment strategy used, UKB participants differ from the general population in their age and sex profile (Sudlow et al 2015). Taking these differences into account using the age and sex distribution of the English UK Biobank cohort in early 2020 and Office for National Statistics estimates of the English population from mid-2018 (https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland), we compared disease incidence in UKB vs. the general population. The absolute number of UKB cases was 72.9% (70.0-79.3%) of the expected total (p=1.7×10^{-13}; Cochran-Mantel-Haenszel test stratified by age and sex), perhaps indicating that UKB participants are healthier or less exposed on average than the general population. Even taking into account this difference, the relative incidence of severe COVID-19 by age and sex was not identical between non-UKB inpatients and UKB inpatients (p=0.011 females, p=0.005 males, chi-squared goodness-of-fit tests). The age distribution of male UKB participants categorised as inpatients and non-inpatients was similar (p=0.83, chi-squared goodness-of-fit test), and although the age distributions of female UKB participants categorised as inpatients and non-inpatients differed

(p=5.0×10^{-5}), visual inspection (Figure 5B, C) did not reveal an obvious trend to these departures, which appear small in magnitude. In conclusion, incidence of severe COVID-19 by age and sex is similar but not identical between the English UKB cohort and the rest of the population, with 27.1% fewer cases than expected.

We investigated the robustness of our method of identifying hospital inpatients, on which we base the definition of severe COVID-19. We compared the proposed definition, stated above, to an alternative that identifies hospital inpatients as only those with Requesting Organisation Type equal to "Hospital Inpatient". This definition is less sensitive, identifying only 316 inpatients compared to 552. Figure 6 indicates that the alternative definition may be more specific, because the odds ratios of severe COVID-19 were larger (further from 1) than under the proposed definition. We conclude that greater sensitivity of the proposed definition trades off some specificity compared to a more stringent alternative, and therefore modestly dilutes the magnitude of age and sex-specific differences in severe COVID-19 incidence. Researchers may wish to investigate this sensitivity-specificity trade-off in downstream analyses.

7.3 Power calculations for genome-wide association studies

If we could predict how numbers of ascertained cases of severe COVID-19 in the UKB cohort will increase over time, we could calculate the power of statistical analyses to discover risk factors. Since it is difficult to predict the outbreak trajectory, we investigated the statistical power to detect human genetic risk factors for severe COVID-19 as a function of the possible number of future cases. This is useful because the absolute number of new cases in UKB can be roughly estimated as a proportion of the total new cases published daily, assuming currently testing trends continue, as detailed above.

We considered the power of a genome-wide association study (GWAS) to detect a rare human allele that increases the risk of severe COVID-19. This is not the only analysis of interest, as UKB contains detailed information on lifestyle and medical variables in addition to human genetics. However, the calculation for a GWAS is instructive because of its large scale (circa 800,000 directly genotyped variants) and standardized approach. In particular, testing on this scale attracts a highly stringent multiple testing significance threshold of p<5×10^{-8}, so the GWAS example is arguably a conservative illustration compared to other analyses.

Figure 7 shows the smallest detectable odds ratio at 80% power, as a function of sample size and risk allele frequency. The odds ratio quantifies the relative probability of case vs control status for individuals possessing the risk vs protective allele. We made a range of simplifying assumptions: that the sample frequency of cases is 72.9% of the population frequency (see above), that the variant is not on a sex chromosome, that the variant is in Hardy-Weinberg equilibrium, that two copies of the risk allele squares the odds ratio, that the white European subset of circa 350,000 individuals is analysed, that population stratification of the risk allele is negligible and that the risk allele is the causal variant, rather than a linked variant. We calculated the power using the bpower function of the Hmisc package in R (https://cran.r-project.org/package=Hmisc).

The calculations indicate that even with 5,000 cases, the above analysis would have high (80%) power to detect only relatively large odds ratios exceeding 1.2 for risk alleles at 10% population frequency. For rarer risk alleles (1% and 0.1% respectively), the odds ratios detectable with

80% power increase to 1.6 and 2.8 respectively. While odds ratios of these magnitudes are known for some infection susceptibility variants, many known variants possess more modest odds ratios below 1.2 (Mozzi et al 2018).

Our calculations do not take into consideration boosts in power that can be achieved by various means, including pooling the effects of multiple variants within or between genes using analyses of various kinds (e.g. Willer and Abecasis 2010) or meta-analysis of the sort planned by the COVID-19 Host Genetics Initiative (www.covid19hg.org), which aims to combine signals across multiple cohorts.

8. Discussion

There are several limitations to this work. We rely on microbiological testing which has been largely restricted to hospitalized cases. Under-ascertainment of severe COVID-19 in community settings, for example nursing homes, is therefore highly likely. Even where SARS-CoV-2 PCR tests have been performed, we cannot assume that the assay is fully sensitive. Since COVID-19 severity scores are not yet readily available, we have made the assumption that hospitalized cases with SARS-CoV-2 positive tests are a proxy for severe COVID-19. The method by which hospital inpatients are identified may affect downstream analyses. Additionally, we have not distinguished those individuals with only positive tests from those with a mixture of positive and negative tests. Integration of further data sources may mitigate some of these limitations, adding information on clinical disease severity and admission to intensive care, which is collected for some individuals in the PHE CHESS and ICNARC databases.

We are unable to assess exposure to SARS-CoV-2 in most UKB participants. This has important implications for case-control studies because we cannot distinguish individuals who have not contracted SARS-CoV-2 following exposure from those who have not been exposed. As the outbreak progresses, exposure and cases of severe COVID-19 will increase. Any case-control definition is thus inherently dynamic, and this will affect analysis and interpretation. Moreover, the nature of the SGSS resource and future changes in national testing mean that interpretation of the data feed remains fluid; we will review such changes and provide updates via the project website www.bugbank.uk.

Despite its limitations, the linkage of COVID-19 test results to the UKB provides a valuable resource to the international research community that has the potential to uncover new risk factors for severe infection. UKB is one of the largest and closest-studied cohorts in the world. A wide range of detailed epidemiological risk factors encompassing lifestyle and medical variables are available to UKB registered researchers to study, in addition to human genotyping and a variety of other technologies such as exome sequencing in some participants. Our study has the potential for wider impact beyond enabling urgent research into COVID-19, because it makes it possible to prospectively sample microbiological cultures from UKB participants that will – subject to detailed assessment through an ongoing culture collection feasibility study – afford an opportunity to study microbiological and molecular genetic risk factors for a range of other important pathogens.

9. Author statements

9.1 Authors and contributors

Conceptualization: NA, DHW, DHW, AMOC
Data curation, software, validation: JA, AMOC
Formal analysis: JA, AMOC
Funding acquisition: DJW
Investigation, Methodology, Visualization, Writing: JA, JKR, NA, DJW, DHW, AMOC
Project administration: NA, DJW, DHW, AMOC
Resources: AMOC
Supervision: DJW, DHW, AMOC

9.2 Conflicts of interest

The author(s) declare that there are no conflicts of interest.

9.3 Funding information

This work was funded by a Big Data Institute Robertson Fellowship (DJW). DJW is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant 101237/Z/13/B).

9.4 Ethical approval

Public Health England gathers data from National Health Service microbiology laboratories, storing it in the SGSS database for epidemiological analysis, an activity permitted under Section 251 of the National Health Service Act 2006, which allows processing of patient data for defined purposes, including public health surveillance. Research in the UK Biobank is conducted under Research Ethics Committee (REC) approval 16/NW/0274. Participants in the UK Biobank have given written, informed, revocable consent for UK Biobank to follow their health using linkage to electronic health-related records.

9.5 Consent for publication

Not required.

9.6 Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 53100. We acknowledge the support of the UK Biobank participants, the UK Biobank staff and members of the National Infection Service Information Management team. We thank Chris Spencer for helpful comments.

10. References

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y. and Yu, T., 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, *395*(10223), pp.507-513.

Chinese Preventive Medicine Association, 2020. An update on the epidemiological characteristics of novel coronavirus pneumonia (COVID-19). *Chin J Epidemiol*, *41*(2), pp.139-144.

Dong E, Wu D and L. Gardner (2020) An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infectious Diseases*, in press doi: 10.1016/S1473-3099(20)30120-1

Hilton, B., Wilson, D. J., O'Connell A.-M., Ironmonger, D., Rudkin, J. K, Allen, N. and D. Wyllie (2020) Microbial isolation in English participants in the UK Biobank cohort: comparison with the general population. *medRxiv* doi:10.1101/2020.03.18.20038281

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. and Cheng, Z., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), pp.497-506.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y. and Xing, X., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* 382: 1199-1207.

Mozzi, A., Pontremoli, C. and Sironi, M., 2018. Genetic susceptibility to infectious diseases: Current status and future perspectives from genome-wide approaches. *Infection, Genetics and Evolution*, *66*: 286-307.

Novel Coronavirus Pneumonia Emergency Response Epidemiology Team (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *China CDC Weekly* **2**(8): 113-122.

Shin, S. K., & Sanders, G. L. (2006). Denormalization strategies for data retrieval from data warehouses. *Decision Support Systems*, *42*(1), 267-282.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. (2015) UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**(3): e1001779.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y. and Zhao, Y., 2020. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *Jama*, *323*(11), pp.1061-1069.

Willer, C.J., Li, Y. and Abecasis, G.R., 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), pp.2190-2191.

Yang, Y., Lu, Q., Liu, M., Wang, Y., Zhang, A., Jalali, N., Dean, N., Longini, I., Halloran, M.E., Xu, B. and Zhang, X., 2020. Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. *medRxiv*.

Yoshikawa TT. (2000) Epidemiology and unique aspects of aging and infectious diseases. *Clinical Infectious Diseases* **30**(6): 931-3.

Zhu, N et al (2020) A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382:727-733.

Figures and tables

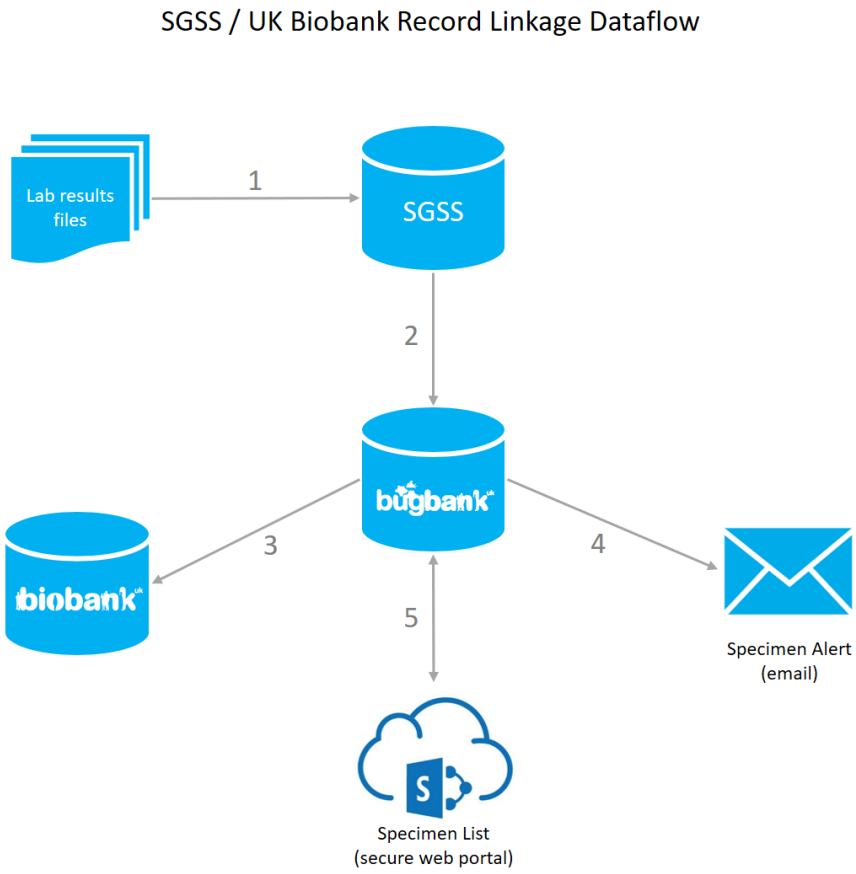

Figure 1. ***Information flow for identifying infection events among UKB participants in SGSS and issuing lab alerts to retrieve those micro-organisms. 1. A laboratory results file is received by SGSS from an NHS or PHE lab. 2. Hourly, an agent checks SGSS for new UKB infection events, and adds any to a separate database in PHE, 'Bugbank'. The agent copies specimen and AMR susceptibility records. 3. Periodically, an extract of the data is transferred securely to UK Biobank for incorporation into their system. 4. Daily, the agent sends an email alert to each active NHS or PHE lab. The email contains minimal information necessary for the lab to retrieve micro-organisms from UKB participant infections. 5. A secure SharePoint site provides a front end to view each lab's specimens in PHE's 'Bugbank' database and records whether each micro-organism has been recovered, is missing, or the record veracity has been questioned by the lab.***

Sent: 10 April 2020 08:45:01 (UTC) Dublin, Edinburgh, Lisbon, London
To: [redacted]
Subject: [ALERT]: New UK Biobank sample(s) to collect

Hello,

3 new UK Biobank samples have arrived at your lab (OXFORD JOHN RADCLIFFE):

Specimen Number	Organism Species	Specimen Date
[redacted]	ESCHERICHIA COLI	2020-04-06
[redacted]	ESCHERICHIA COLI	2020-04-06
[redacted]	ESCHERICHIA COLI	2020-04-05

Additionally there remain 1845 unrecovered specimens from yesterday's alert.
Where applicable, please mark any collected samples as "Collected" and any noncollectable samples as "Missing".

Thank you,
Bugbank Alert Bot

Figure 2. Example email alert to retrieve micro-organisms cultured from UKB participants' infections. The alert is sent automatically from PHE Colindale to the NHS or PHE lab. It contains minimal information necessary to retrieve the micro-organisms.

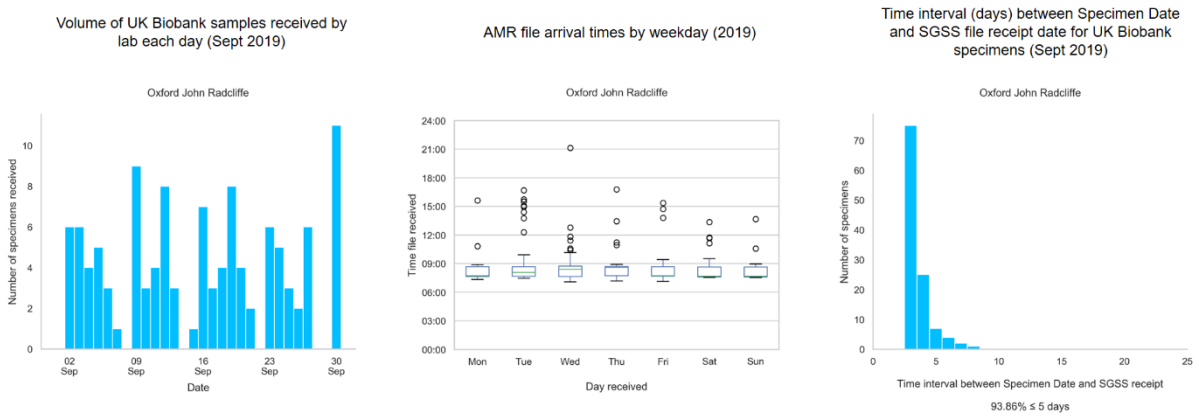

Figure 3. Summary of UKB infection events at the John Radcliffe Hospital, Oxford during September 2019. The summary indicates the volume of events by date (left), the time of day (middle), and the time to record receipt in SGSS (right).

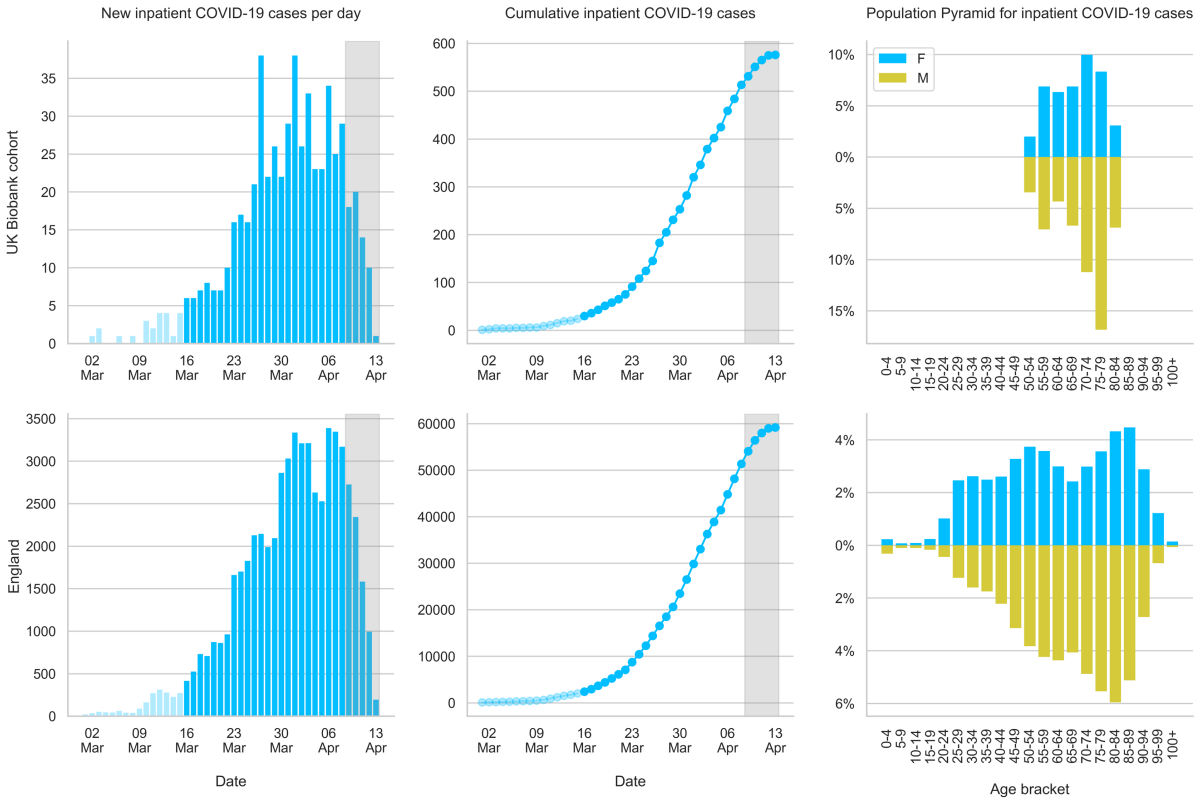

Figure 4. Demographic features of individuals with COVID-19 among English UK Biobank participants (top) and all individuals in the Public Health England Second Generation Surveillance System (bottom). Only hospital inpatients are shown since these cases can be inferred as severe COVID-19 because only severe cases were admitted to hospital from 16th March onwards. COVID-19 is determined by positive PCR for SARS-CoV-19. Panels show total number of new cases per day (left), cumulative number of cases (middle) and the age and sex distribution of cases (right). The dark grey shaded region (left and middle panels) highlights the reporting lag period for some cases, assessed as around 5 days.

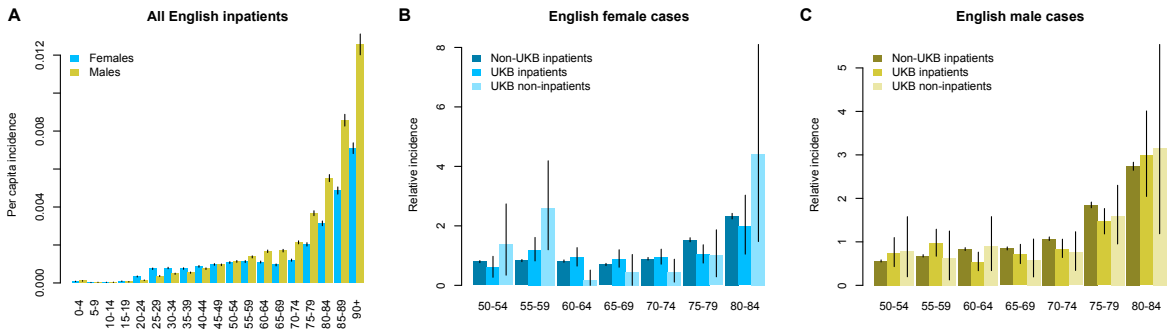

Figure 5. Incidence of SARS-CoV-2 positive individuals in England, 16th March – 10th April. Per capita incidence is show for all English inpatients (A). Relative incidence is compared between non-UKB inpatients, UKB inpatients and UKB non-inpatients for females (B) and males (C). Vertical black lines indicate 95% confidence intervals calculated assuming a Poisson distribution for the underlying counts. Incidence was calculated using the known age and sex distribution for England as a whole, and English UKB participants.

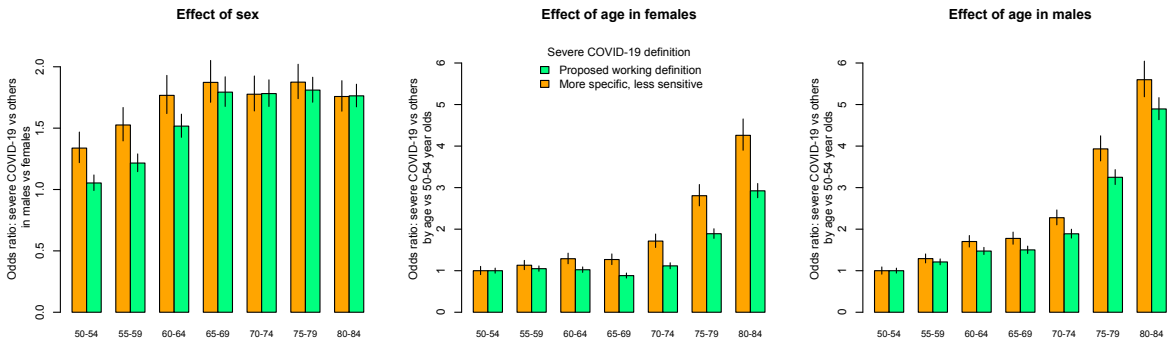

Figure 6. Impact of alternative definitions of severe COVID-19 on estimated effects of age and sex on incidence. PCR-positive hospitalized inpatients are considered to represent severe cases. Two methods of identifying inpatients are compared, the proposed working definition (green) and a more specific but less sensitive method (orange). For each definition, the effects of age and sex on the odds of severe COVID-19 were estimated using fisher.test in R and compared.

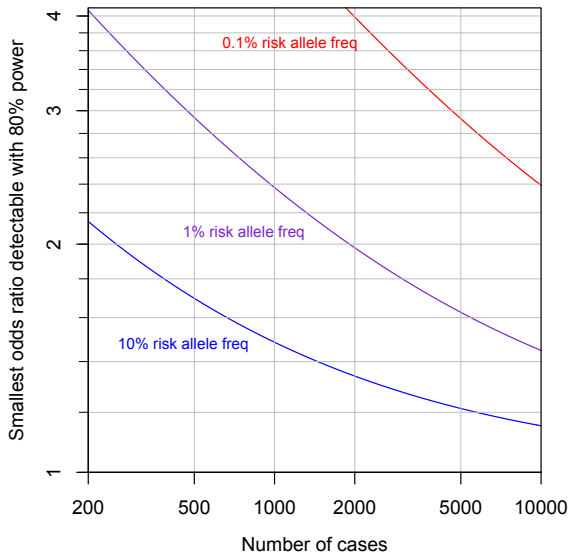

Figure 7. Power calculations for genome-wide association studies. The smallest odds ratio (case/control status vs risk/protective allele) detectable with 80% power is shown as the number of cases increases from 200 to 10,000, assuming a genome-wide significance threshold of 5×10^{-8}.