Tan, Y., & Timpson, N. J. (2022). The UK Biobank: A Shining Example of Genome-Wide Association Study Science with the Power to Detect the Murky Complications of Real-World Epidemiology. *Annual Review of Genomics and Human Genetics*, *23*, 13.1-13.21. https://doi.org/10.1146/annurev-genom-121321-093606

UKBiobank – a shining example of genomewide association study science with the power to detect murky complications in real-world epidemiology

Vanessa Y. Tan[1,2], Nicholas J. Timpson[1,2]
[1]Medical Research Council (MRC) Integrative Epidemiology Unit, University of Bristol, Bristol United Kingdom
[2]Bristol Medical School, University of Bristol, Bristol, United Kingdom

Corresponding author: Professor Nicholas J. Timpson, Address: MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, email: N.J.Timpson@bristol.ac.uk

Abstract
Genome-wide association studies (GWAS) studies have successfully identified thousands of genetic variants reliably associated with human traits. Albeit restricted to certain variant frequencies, this has led to an improvement in our understanding of the genetic architecture of complex traits and diseases. The availability of large genomic biobanks such as the UK Biobank (UKBB) study have brought substantial analytical power to association studies. The dramatic expansion of GWAS sample sizes improves power of estimation of effect sizes, genomic prediction and the potential for applied analyses such as those relating to causal inference. However, in the same moment, the availability of substantial analytical power and enabling analytical capacity can increase the complications and inferential complexity associated with GWAS and other applied analyses. In this review, we discuss the revolutionary impact that UKBB has had in the GWAS era and some of the opportunities and challenges of using data from this world-leading study.

Introduction
Genetic architecture is defined by the genetic variants influencing a trait or disease outcome and is dependent characterised by the number of genetic variants, their effect size, their allele frequency and the possible interactions with each other and environmental factors[1]. Uncovering the genetic architecture of a complex traits or disease is central to understanding what underpins observed trait variation and potentially helps further work aimed at dissecting this. Various techniques are useful in the task of assessing genetic architecture and in the last decade, technological advances have enabled measurement of genetic variation at hundreds of thousands of markers across the human genome. Genome-wide association studies (GWAS) studies have exploited these developments and successfully identified thousands of genetic variants reliably associated with human traits. Albeit restricted to certain variant frequencies, this has led to an improvement in our understanding of the genetic architecture of complex traits and diseases[2]. Early successful applications of GWAS have primed the rapid growth and increasing availability of large genomic biobanks such as the UK Biobank (UKBB) study[3,4], which have brought substantial analytical power to association studies. Here we discuss the revolutionary impact that UKBB has had in the GWAS era and some of the opportunities and challenges of using data from this world-leading study.

The revolutionary impact of UKBB in the GWAS world
UKBB is a prospective cohort study that recruited half a million individuals (40 to 69 years of age) across the United Kingdom between 2006 and 2010[3]. The study is a large-scale biomedical resource integrating genome-wide genetic data with deep phenotype data- including data from lifestyle questionnaires, physical measures, biomarkers in blood and urine, accelerometry and multimodal imaging. The unprecedented size of the UKBB cohort, together with the extensive phenotyping and genome-wide genotype data (supplemented with high-density imputation) has enhanced power for genetic discovery and enabled well-powered GWAS of hundreds of quantitative traits, including anthropometric traits[5], blood traits[6], cognitive traits[7] and numerous blood and urine biomarkers[8] to be conducted. Further

to this enormous collection of data, the access infrastructure provided with the UKBB study has led to UKBB being one of the most enabling human genetics bioresources ever generated. To-date, 514 peer-reviewed GWAS publications have been published from this one resource and it is often the cases that UKBB is the primary data provider in analyses undertaken.

In the context of the capacity and coverage of GWAS arrays (and imputation), sample size is one of a series of key determinants in the deployment of GWAS for the discovery of genetic loci associated with complex traits[2]. This has been evident since the first application of a-hypothetical approaches to gene variant analysis[9–11] and – dependent on the composition of heritable contributions – the discovery of new loci tends has increased in an almost linear fashion with increasing sample size[12]. The availability of genome-wide genotype data collected from all participants, together with the vast amount of phenotype data available in UKBB has generated a singular resource of considerable size that provides opportunities for the discovery of new genetic associations and the genetic basis of complex traits and diseases[4]. The gain in power in the UKBB has been exemplified in the most recent height and BMI GWAS meta-analysis[5] which combined results from a single large GWAS conducted using data from UKBB, with previously published GWAS of height and BMI conducted by the GIANT study. With the increased sample size, the number of genomic loci associated with height and BMI was increased compared to the previously published height and BMI GWAS with improved accuracy of genetic predictors from SNPs at these loci[5]. The near independent SNPs explained approximately 24.6% and 6% of the variance in height and BMI, respectively, representing approximately a 1.9 and a 3.2-fold improvement in comparison with previous BMI GWAS meta-analysis without including UKBB[5].

The combination of large scale and extensive phenotypic and genotypic data from UKBB enables the rigorous investigation of the genetic basis of diseases, not just through sample size alone, but also through phenotypic precision[13]. Although expanding collections of genotype and phenotype data from non-UKBB studies have provided a boost in statistical power, research has been hindered by measurement differences, inaccurate phenotypic measurements and genuine disease heterogeneity. A challenge to GWAS studies is the ability to combine genetic data with phenotypic precision and hence to both enhance analytical power technically and to tighten the focus of downstream interpretation of findings around pertinent association signals[14,15]. In UKBB, the presence of a combination of deep phenotypic data collection and scale has generated both biological insight and extensive records of genotype/phenotype association. A good example of this can be found in the use of detailed brain imaging data to examine the genetic architecture of brain structure and wich undertook GWAS of >3000 functional and structural brain imaging phenotypes in >8000 UKBB participants. This work was not only able to show that many of these phenotypes are likely to be heritable, but that they lie in phenotypic clusters showing reliable genetic associations[16]. This has been further exemplified by Aragam et al[17] which used data from UKBB to perform a GWAS for heart failure and found that phenotypic refinement of all-cause heart failure facilitates the discovery of novel genetic signals that reflect distinct etiologic heart failure subtypes.

Overlapping phenotypes and harnessing the phenome

Most GWAS studies only analyse a single trait; however, these do not exploit information of summary statistics from GWAS of other correlated traits. Joint association analysis of multiple traits in GWAS studies offers several advantages compared to single trait analyses and have been a well-used approach in the undertaking of GWAS in UKBB given the nature of the resource. Firstly, multivariate analysis can boost statistical power as it takes into account cross-trait covariance of genetically correlated traits which is often ignored in univariate analyses[18,19]. Secondly, as multivariate methods tests associations with a set of traits using a single test, the multiple testing burden is reduced due to the reduced number of tests performed[18,20]. Lastly, where a single genetic variant is highly pleiotropic and

associated with multiple traits, multivariate GWAS analysis is more consistent with the biology of the traits compared to univariate analyses[21]. Several multi-trait techniques have been developed[22–27] to conduct joint analysis of multiple traits. For example, Multi-trait analysis of GWAS (MTAG) has been developed which allows the joint analysis of multiple traits in population-based GWAS, therefore increasing statistical power to detect genetic associations for each analysed trait. AS good example of this type of analysis, using MTAG, a recent study conducted a joint GWAS analysis of four hearing related traits from UKBB and identified 31 new risk loci for hearing difficulty[28].

Applied analyses enabled by a GWAS backbone

One of the most intuitive applied analyses built on the success of well undertaken and powered GWAS studies is the examination of polygenic risk score (PRS) analysis. A genomewide PRS integrates all available common variants associated with the trait from the largest or most informative GWAS into a single quantitative measure of inherited susceptibility. Several studies have limited success in obtaining meaningful predictive power[29,30]. However, previous effects to create an effective polygenic score were limited by three challenges[31]: 1) small sample size of GWAS study, which affected the precision of the estimated impact of individual variants on trait; 2) limited computational methods for creating the PRSs; and 3) lack of large datasets to validate and test PRS. To overcome some of these challenges, a recent study by Khera et al[32] used data from UKBB as a validation dataset to test the ability of the BMI PRS to predict measured BMI. The study demonstrated the ability to use PRS to identify individuals at greatest risk of obesity with over 40% of individuals achieving a PRS score in the top decile found to be obese compared to 10% of individual in the bottom decile[32].

Mendelian randomization (MR) is an analytic technique that uses genetic variants as instruments to estimate the causal effect of an exposure on an outcome of interest[33]. By exploiting the properties of genetic data, MR analyses provide an alternative source of evidence when estimating causal effects and attempting to minimise limitations through confounding, bias and reverse causation. MR analyses can be undertaken using individual studies with exposures, outcomes and genetic data, but also using the results from existing GWAS studies[34,35]. MR-base[36] is an established and freely accessible, online platform, which combines a database of genome-wide association study results. This together with an interface for performing MR and sensitivity analyses, has simplified the implementation of MR studies and enabled users to explore millions of potentially causal associations. The expansion of large-scale GWAS using data from UKBB rapidly expanded the collection of genetic variants reliably associated with human characteristics and health conditions. For example, since the release of genome-wide association data, GWAS has been conducted for thousands of phenotypes by the Neale lab[37]. These data has been incorporated in the IEU OpenGWAS database[38], an open source, open access, scalable and high-performance cloud-based data infrastructure that imports and publishes complete GWAS summary datasets and meta-data for the scientific community. Taken together, this open GWAS data resource and the development of results based MR application and newly available analytical tools has enabled causal inference analysis[38]. For example, a recent MR study found evidence that fat mass exerted detrimental effects on most cardiometabolic traits by using the IEU OpenGWAS database[38] to obtain genetic instruments for body composition measures from GWAS conducted in UK Biobank[39].

A different, but similarly applied approach using GWAS analysis and results, is that of the examination of shared genetic architecture or genetic correlation. Genetic correlation is a quantitative parameter which quantifies the shared heritable contribution to two traits. Identifying genetic correlations between complex traits and diseases can provide useful insights into disease etiology, can help identify potential causal relationships[40] and increase

understanding as to shared biological contributions to apparently independent traits. Methodological approaches that estimate SNP-based heritability and genetic correlations from genome-wide association studies, such as LD score regression[40] (LDSR) have proven to be powerful tools to provide a robust estimate of the genetic correlations between different traits and diseases and for helping to dissect the genetic architecture of common traits and diseases. LDSR relies on GWAS summary statistics and is not biased by sample overlap and thus is invaluable in increasing our knowledge of the genetic contribution to complex traits. A major challenge preventing accurate estimation of genetic correlation is that GWAS with small effective sample sizes have insufficient power for LSDR to detect polygenic effects, leading to near-zero estimates of heritability. Recently, data from UKBB has been used to accurately estimate the SNP-heritability of 22 complex traits and disease traits[41] and the genetic correlation between various traits and diseases[42]. The LDSR analysis of more than 2000 phenotypes in UKBB found substantial variance explained by common SNPs for a broad range of human traits and diseases.

Linkage – the availability of EHR data and records

Given the prospective nature of the UKBB study, a key strength of the cohort is the collection of data, biosamples and exposures at baseline which can be linked to electronic health records (EHR) for prospective followup. These resources include death and cancer registries, primary and secondary care records, and there is potential to follow-up the health of all participants over-time in the absence of attrition. This linkage to health outcome data provides opportunities to conduct research on common diseases such as ischemic heart disease and various cancers and to further expand the portfolio of GWAS studies embedded in the UKBB resource. This powerful design also enables conditions that are difficult to study retrospectively including dementia and rapidly fatal conditions such as pancreatic or lung cancer. For example, a recent pan-cancer GWAS provided insights into complex genetic architecture of cross-cancer susceptibility by using linked cancer registry data in UKBB and the Kaiser Permanente Genetic Epidemiology Research on Adult Health and Aging cohort[43].

Research practice, sharing and the "democratisation of data"

UKBB is an open-access resource that encourages researchers from around the world- including those from the academia and industry- to access the data and biological samples for any health-related research that is in the public interest (www.ukbiobank.ac.uk). The open-access nature of the UKBB study promotes innovative science by enabling international scientists to apply for the data quickly and easily through an application process to benefit from this vast resource[44]. Recently, to accommodate the vast scale of the UKBB resource, the UKBB has launched a unique and innovative Research Analysis Platform (RAP), a cloud-based system that allows streamlined access to approved researchers from anywhere in the world and to enable data to be analysed easily and cost-effectively as the resource grows in complexity and scale. The open access nature of the data has also encouraged collaborations with large international consortia, resulting in the rapid advancements in dissecting the genetic architecture of complex traits which would not have been possible with under-powered studies.

The reality of GWAS – power, polygenicity, sampling frame and interpretation

UKBB is an outstanding example of the value that can be achieved from large sample size combined with genetics, extensive and deep phenotyping and linkage to health records. The gain in power in the UKBB cohort is clear and has led to an increase in loci discovery in GWAS studies, in particular for loci that are less common and/or with smaller effects[2]. For example, the first BMI GWAS (n~5000) identified only genetic variants in the *FTO* locus with relatively large effects on BMI (0.35 kg/m2 per allele)[45,46]. In contrast, the most recent GWAS for BMI which used data from UKBB and the GIANT consortium (n~800,000) identified more than 750 loci, with much smaller effects on BMI (0.04kg/m2 per allele)[5]. Notably, UKBB represented 64.3% of this overall sample size. In this type of work, it is clear that the unprecedented size of the UKBB have provided immense opportunities; however, also can

generate analytical challenges. To focus on two of these, we will examine the potential for population stratification/sub-structure to be important in the presence of specific GWAS studies undertaken at scale and on the potential for detectable phenotypic overlap to complicate downstream interpretation and analysis.

Along with the potential to underrepresent specific groups and reduce generalisability[47], self-selection of participants contributing to the UKBB cohort creates structure within genetic data which has the potential to bias associations and complicate their interpretation. Although UKBB was designed to be representative of the general population of the United Kingdom, the sampling population is volunteer-based and is not representative of the UK population[48] by demographic characteristics. Ultimately, UKBB is a highly-selected sample of the UK population (having a response rate of 5.5%)[49]. This has been illustrated in work by Haworth et al[50] which showed that single genetic variants and genetic scores composed of multiple variants are associated with birth location within UKBB and that geographic structure in genotype data cannot be accounted for using routine adjustment for study centre and genetic principal components. The study also demonstrated that major health outcomes appear geographically structured and that coincident structure in health outcome and genotype data can yield biased associations.

As described above, MR is an analytic technique that has been performed to estimate causal relationship between risk factors and exposures and here serves as a good illustration of the possible complications of analytical power and structure. Population stratification can essentially be thought of as the reintroduction of confounding of the genetic instrument (used to proxy the exposure) and disease outcome; therefore violating the MR assumption that there is no confounding between the genetic instrument and the outcome. Therefore, a GWAS that do not fully account for any ancestral population structure can lead to population stratification[51] and estimates from MR analyses based on the results of that GWAS could potentially be biased by the coincidence of association between genotype, population sub-structure and health. The key question, therefore, is how pervasive this type of structure in data is present and whether it can be demonstrated (as in Haworth et al[50])? The recent GWAS of Coronavirus Disease 2019 (COVID-19) outcomes – substantively aided by UKBB data – is a real-time exemplar of just such potential complication. COVID-19 is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) which has had a profound impact on the health and lives of people worldwide[52]. UKBB has been uniquely positioned to contribute to research into the COVID-19 pandemic. For example, the entire UKBB cohort (around 500,000 participants) had been invited to receive a self-test kit to find out if they have presence of the SARS-CoV-2 antibodies due to past infection (rather than vaccination). UKBB is one of the largest contributors to an international consortium, The Covid-19 Host Genetics Initiative[53] (HGI; http://www.covid19hg.org), that has brought together investigators from around the world to investigate the genetic determinants of COVID-19 susceptibility, as well as severity and outcome. A GWAS study conducted by the HGI initiative have identified at least 15 genome-wide significant loci associated with increased susceptibility and severity of COVID-19, including variants in/near several immune genes and the ABO locus determining ABO blood groups. Many of these loci overlap with previously reported associations with lung-related phenotypes or autoimmune/inflammatory diseases; although some loci have no obvious candidate gene[54]. Such discoveries not only contribute to global knowledge of the biology of SARS-CoV-2 infection but provide the genetic evidence for drug targets and drug repurposing and help in the development of genetically informed risk assessment of COVID-19 susceptibility. The publicly available GWAS results for COVID-19 susceptibility and severity has also enabled Mendelian randomization (MR) studies to be conducted to evaluate the causal effect of various exposures on COVID-19 outcomes. Indeed, as of December 2021, there were 60 Mendelian randomization studies that have been conducted (**Table 1**).

In an effort to shed light on potential complexities involved in the application of new GWAS results in this way, we evaluated the implications of population structure for GWAS of COVID -19 susceptibility and severity in UKBB. We did this by considering the association between the polygenic scores for COVID-19 susceptibility and severity with birth location (Supplementary Methods). For COVID-19 susceptibility, a polygenic risk score representing the aggregate estimated common genetic contributions to these outcomes was associated with birth location in the model that adjusting for genotyping array and study centre (**Figure 1**). These associations were attenuated in models incorporating adjustment for 40 genetic principal components which were able to capture structure in the genetic data available, however this was not always the case for all COVID-19 outcomes. In contrast, for COVID-19 severity (case-only analysis), associations between polygenic risk and location were actually more pronounced in models incorporating adjustment for 40 genetic principal components, potentially reflecting the potentially biasing effects of association analysis within stratified samples; in this case COVID-19 only participants.

Outside of the new challenge presented by SARS-CoV-2 and COVID-19, a second exemplar lies in the notion that one cannot assume an increased ability to deconstruct networks of complex biological association with the availability of big omics and big GWAS; indeed, whilst there will be an ability to discover genetic association signals with good analytical power, redundancy and complexity can interrupt direct interpretation. A good example of this can be found in the analysis of high throughput metabolomic data. Metabolomic profiles are the result of genetic and non-genetic factors and provide a read-out of biological processes and can functionally link genetic loci to disease risk factors and disease outcomes[55–57]. Metabolomics technologies based on mass spectrometry (MS) and nuclear magnetic resonance (NMR) have enabled the systematic quantification of hundreds of metabolites (the 'metabolome') from a single biological sample. The analysis of metabolites has enabled a more thorough exploration of an individual's metabolic status, offering new opportunities to improve our understanding of the molecular mechanisms underlying human traits and diseases[58]. Over the last decade, several metabolite GWASs have been performed[59–68] to characterise the genetic architecture of blood metabolite variation and provided an estimation of the heritability of multiple metabolites and provided insights into the biological and clinical relevance of these genetic associations[64,69]. Recently, metabolic biomarker data quantified using NMR in approximately 121,000 participants have been made available from UKBB[70,71]. The availability of this large-scale omics measurement combined with genome-wide data has maximised the power to discover genetic loci for a given metabolite and to provide a better understanding of the genetic architecture of blood metabolites.

However, the genetic architecture of blood metabolites is complicated by the high correlation structure and shared biology of the metabolites which causes complexities when analysing the causal association between individual metabolites and disease outcomes using MR analyses. This was exemplified in a recent study which demonstrated that genetic instruments associated with metabolites were likely to be highly pleiotropic, with few SNPs found to be associated with specific metabolites[72]. Furthermore, there was high degree of pleiotropy for metabolite-associated SNPs with modifiable risk factors and other disease endpoints. As most metabolites have only a small number of instruments, statistical methods aiming to correct for these biases (e.g MR-Egger[73] and MR-PRESSO[74]) is not possible, nor is the use of techniques designed to evaluate the effect of multiple correlated exposures (e.g multivariable MR[75,76]).

To explore this type of complexity further, we undertook a simple analysis here seeking to demonstrate the number of metabolomic features associated with genetic variants at a predefined and stringent threshold in UKBB (supplementary methods). Using recently available NMR data within UKBB and by undertaking a basic GWAS for circulating metabolites, we found that few of the plentiful collection of potential genetic associations which would satisfy conditions to be used as "instruments" within MR analyses (i.e, i) genetic

variant is associated with the exposure; ii) no association between genetic variant and outcome; and iii) genetic variant is independent of any measured or unmeasured confounding factors), few were associated with a specific metabolite. In contrast, we observed that numerous loci showed high level of multi-metabolite association with a median of 34 metabolites associated with each loci (**Figure 2**). The profound overlapping of association signals across metabolites is clearly a complicating factor and one which would potentially violate assumptions made in analyses such as MR. That is not to say that these associations are neither uninformative nor are these issues insurmountable, rather that they are clear markers of the potential issues that need to be considered when power and precision are able to generate strong association profiles. For studies investigating whether metabolites could be biological pathways relevant to disease onset, a potential way to solve this problem is to conduct profile comparison analysis to examine the overlap between the metabolomic profile of prospective disease risk with that of the risk factor of the disease (e.g BMI) to identify biological pathways relevant to disease onset[72].

Discussion

UKBB is a shining example of the impact of large, open-access population biobanks in increasing the power to understand the genetic architecture of common traits and diseases. Amongst a wider set of potential benefits not all considered here, the dramatic expansion of GWAS sample sizes improves power of estimation of effect sizes, genomic prediction and the potential for applied analyses such as those relating to causal inference. However, in the same moment, the availability of substantial analytical power and enabling analytical capacity can increase the complications and inferential complexity associated with any one specific analysis. For example – as described here and previous studies[50,77] – population structure exists within the UKBB and may have the potential to bias association results or their interpretation. The presence of population structure is challenging, requiring methods that are specific to the analytical context and trait. If not properly corrected for, the sampling structure can generate properties in data that can lead to biased inference. Caution is therefore needed in the interpretation of GWAS results using data from UKBB, particularly for loci which demonstrates strong residual associations with birth location, even after adjustment for population stratification[77].

Despite this, and other limitations mentioned here or elsewhere, UKBB remains an extraordinary resource. Measured by data, output, enabling capacity or likely future contribution, the resource has undeniably shaped the modern era of GWAS. Most of the problems noted in the analysis of results from UKBB are, and likely will be, the result of mis-interpretation of results generated from the UKBB sampling frame – not from the sampling frame itself. Used for appropriate analyses and with results interpreted in the context of the specific nature of the sample that is UKBB, there is no doubt that UKBB will continue to be a shining light in the field of human GWAS.

References

1. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* 2018.
2. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 2017.
3. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*. 2015;
4. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;
5. Yengo L, Sidorenko J, Kemper KE, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Hum Mol Genet*. 2018;

6. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. 2016;
7. Hill WD, Marioni RE, Maghzian O, et al. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol Psychiatry*. 2019;
8. Sinnott-Armstrong N, Tanigawa Y, Amar D, et al. Genetics of 38 blood and urine biomarkers in the UK Biobank. *bioRxiv*. 2019;
9. Spencer CCA, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*. 2009;
10. Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;
11. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* 2008.
12. Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JPA. The power of meta-analysis in genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* 2013.
13. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: Opportunities for cardiovascular research. *Eur. Heart J.* 2019.
14. Smith JG. Molecular Epidemiology of Heart Failure: Translational Challenges and Opportunities. *JACC Basic to Transl. Sci.* 2017.
15. Sluis S van der, Verhage M, Posthuma D, Dolan C V. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PLoS One*. 2010;
16. Elliott LT, Sharp K, Alfaro-Almagro F, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*. 2018;
17. Aragam KG, Chaffin M, Levinson RT, et al. Phenotypic Refinement of Heart Failure in a National Biobank Facilitates Genetic Discovery. *Circulation*. 2019;
18. Zhu W, Zhang H. Why do we test multiple traits in genetic association studies? *J Korean Stat Soc*. 2009;
19. Allison DB, Thiel B, Jean P St., Elston RC, Infante MC, Schork NJ. Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages. *Am J Hum Genet*. 1998;
20. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol*. 2008;
21. Chavali S, Barrenas F, Kanduri K, Benson M. Network properties of human disease genes with pleiotropic effects. *BMC Syst Biol*. 2010;
22. Galesloot TE, Steen K Van, Kiemeney LALM, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. *PLoS One*. 2014;
23. Porter HF, O'Reilly PF. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci Rep*. 2017;
24. Maier R, Moser G, Chen GB, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet*. 2015;
25. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet*. 2017;
26. Baselmans BML, Jansen R, Ip HF, et al. Multivariate Genome-wide and integrated transcriptome and epigenome-wide analyses of the Well-being spectrum. *bioRxiv*. 2017;
27. Turley P, Walters RK, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet*. 2018;
28. Kalra G, Milon B, Casella AM, et al. Biological insights from multi-omic analysis of 31 genomic risk loci for adult hearing difficulty. *PLoS Genet*. 2020;
29. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: Case-control and prospective cohort analyses. *Lancet*. 2010;

30. Loos RJF, Janssens ACJW. Predicting Polygenic Obesity Using Genetic Information. *Cell Metab.* 2017.
31. Khera A V., Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 2018.
32. Khera A V., Chaffin M, Wade KH, et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*. 2019;
33. Smith GD, Ebrahim S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 2003.
34. Lawlor DA. Commentary: Two-sample Mendelian randomization: Opportunities and challenges. *Int. J. Epidemiol.* 2016.
35. Zhao Q, Wang J, Spiller W, Bowden J, Small DS. Two-sample instrumental variable analyses using heterogeneous samples. *Stat Sci*. 2019;
36. Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic causal inference across the human phenome. *Elife*. 2018;
37. Neale BM. UK Biobank-Neale lab [Internet]. 2018. Available from: http://www.nealelab.is/uk-biobank
38. Elsworth B, Lyon M, Alexander T, et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*. 2020;
39. Zeng H, Lin C, Wang S, Zheng Y, Gao X. Genetically predicted body composition in relation to cardiometabolic traits: a Mendelian randomization study. *Eur J Epidemiol*. 2021;
40. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet*. 2015;
41. Hou K, Burch KS, Majumdar A, et al. Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat Genet*. 2019;
42. Neale BM. Genetic correlation between traits and disorders in the UK Biobank [Internet]. 2020. Available from: https://ukbb-rg.hail.is
43. Rashkin SR, Graff RE, Kachuri L, et al. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun*. 2020;
44. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. *J. Intern. Med.* 2019.
45. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (80-)*. 2007;
46. Scuteri A, Sanna S, Chen WM, et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet*. 2007;
47. Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* 2019.
48. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *Am J Epidemiol*. 2017;
49. Swanson JM. The UK Biobank and selection bias. *Lancet* 2012.
50. Haworth S, Mitchell R, Corbin L, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun*. 2019;
51. Lawson DJ, Davies NM, Haworth S, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* 2020.
52. Sharma A, Tiwari S, Deb MK, Marty JL. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies. *Int J Antimicrob Agents*. 2020;
53. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet*. 2020;

54. Mapping the human genetic architecture of COVID-19. *Nature*. 2021;
55. Lewis GD, Wei R, Liu E, et al. Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. *J Clin Invest*. 2008;
56. Blasco H, Nadal-Desbarats L, Pradat PF, et al. Biomarkers in amyotrophic lateral sclerosis: Combining metabolomic and clinical parameters to define disease progression. *Eur J Neurol*. 2016;
57. Yazdani A, Yazdani A, Saniei A, Boerwinkle E. A causal network analysis in an observational study identifies metabolomics pathways influencing plasma triglyceride levels. *Metabolomics*. 2016;
58. Kastenmüller G, Raffler J, Gieger C, Suhre K. Genetics of human metabolism: An update. *Hum. Mol. Genet.* 2015.
59. Kettunen J, Demirkan A, Würtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun*. 2016;
60. Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014;
61. Long T, Hicks M, Yu HC, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet*. 2017;
62. Draisma HHM, Pool R, Kobl M, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun*. 2015;
63. Illig T, Gieger C, Zhai G, et al. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*. 2010;
64. Suhre K, Shin SY, Petersen AK, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. 2011;
65. Rhee EP, Ho JE, Chen MH, et al. A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab*. 2013;
66. Gallois A, Mefford J, Ko A, et al. A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat Commun*. 2019;
67. Rhee EP, Yang Q, Yu B, et al. An exome array study of the plasma metabolome. *Nat Commun*. 2016;
68. Lotta LA, Pietzner M, Stewart ID, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat Genet*. 2021;
69. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet*. 2015;
70. Ritchie SC, Surendran P, Karthikeyan S, et al. Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *medRxiv*. 2021;
71. Julkunen H, Cichońska A, Slagboom PE, Würtz P. Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. *Elife*. 2021;
72. Guida F, Tan VY, Corbin LJ, et al. The blood metabolome of incident kidney cancer: A case-control study nested within the MetKid consortium. *PLoS Med*. 2021;
73. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015;
74. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*. 2018;
75. Sanderson E, Spiller W, Bowden J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat Med*. 2021;
76. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary

data settings. *Int J Epidemiol*. 2019;
77. Cook JP, Mahajan A, Morris AP. Fine-scale population structure in the UK biobank: Implications for genome-wide association studies. *Hum Mol Genet*. 2020;
78. Mitchell R, Hemani G, Dudding T, Corbin L, Harrison S PL. UK Biobank Genetic Data: MRC-IEU Quality Control, version 2 [Internet]. 2019. Available from: https://doi.org/10.5523/bris.1ovaau5sxunp2cv8rcy88688v
79. Wood SN. Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. *Gen Addit Model An Introd with R, Second Ed*. 2020;
80. R Core Team (2020). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria* 2020.
81. Würtz P, Kangas AJ, Soininen P, Lawlor DA, Davey Smith G, Ala-Korpela M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *Am J Epidemiol*. 2017;

Supplementary methods

Implications of population structure for GWAS of Covid-19 susceptibility and severity in the UK Biobank

The UK Biobank study assessment centre sites targeted densely populated areas of England, Scotland and Wales, where a large eligible population could attend in-person assessment with a journey of less than 10 miles[3]. Participants gave informed consent, and the UK Biobank was approved by the North West Multi-centre Research Ethics Committee. This research was conducted using the UK Biobank Resource application 15825, and complied with all relevant ethical regulations.

The assessment centre at which a participant consented was assigned a numerical code (field 54 in the UK Biobank data). In analyses adjusted for assessment centre, these codes were treated as factor variables. Participants who were born in the UK were asked to name their place of birth during a verbal interview at study assessment centres. These answers were used to derive approximate North and East co-ordinates (rounded values, recorded on a metre grid scale from an origin South-West of the UK, fields 129 and 130 in the UK Biobank data). Values less than zero were coded as missing for both variables.

We used the UK Biobank 500K (July 2017) genotype release, for which pre-imputation quality control, phasing and imputation are described elsewhere[4]. Following imputation, we removed variants that were not present within the haplotype reference consortium (HRC) imputation panel and applied a graded filtering on imputation quality. Rarer variants were required to have a higher imputation INFO score (Info>0.3 for minor allele frequency (MAF) >3%; Info>0.6 for MAF 1-3%; Info>0.8 for MAF 0.5-1% and Info>0.9 for MAF 0.1-0.5%). 378 individuals were removed as a result of mismatches between genetic sex and reported sex and 352 individuals with a putative sex chromosome aneuploidy. We performed analysis within individuals who self-reported as “British” and had similar ancestral background from genetic Principal components (PCs) (409,703). We applied an exclusion list containing 79,448 individuals, whilst preferentially removing individuals related to the greatest number of other individuals so that no related pairs remained in the final sample used for analysis. A comprehensive description of quality control methods has been published online[78].

Genetic PCs were supplied by UK Biobank (field 22009). These were calculated using a set of 407,219 unrelated, high-quality samples and 147,604 high confidence markers after pruning for linkage disequilibrium. Participants with missing PCs were excluded from analysis.

Independent genetic variants ($p<5e^{-08}$) associated with COVID-19 susceptibility and severity were taken from the latest Covid-19 GWAS[54] (release 6). Effect allele dosage were extracted from these variants from the filtered UK Biobank genotype data. Effect allele dosage was weighted by reported genetic effect (beta) and then summarised across all the contributing variants to create per-individual PRS.

The relationship between Covid-19 PRS and geographical parameters were modelled using the 'mgcv' package[79] (version 1.8) in R (version 4.0.4)[80]. Traits were modelled against a spline function for either birth northings or birth eastings with minimum adjustment for genotyping array, in the form PRS ~ s(location) + array. Fully adjusted models included factors variables for study centre and 40 genetic PCs. Approximate statistical significance for non-linear terms were taken from the model summary, which estimates a suitable number of degrees of freedom from cross validation.

Pleiotropy of metabolite instruments

We obtained the GWAS summary data for NMR metabolite GWAS (under the batch name 'met-d' conducted using data from UK Biobank via the OpenGWAS database[38]. Briefly, metabolites were measured in a random subset of non-fasting baseline plasma samples (aliquot 3) from 118,466 UK Biobank participants and 1298 repeat-visit samples using high-throughput MR spectroscopy (Nightingale Health Plc; biomarker quantification version 2020). This platform provides simultaneous quantification of 249 metabolic measures including routine lipids, lipoprotein subclass profiling with lipid concentrations within 14 subclasses, fatty acid composition, and various low-molecular weight metabolites such as amino acids, ketone bodies, and glycolysis-related metabolites. More information: http://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220. Technical details and epidemiological applications of Nightingale platform have been previously reviewed[69,81]. Metabolites were transformed using rank-based inverse normal transformation (INT) prior to analyses. As the GWAS analysis was performed on pre-released dataset, there were addition QC filters that were applied that are automatically performed for all general release datasets. In particular, observations where biomarkers were tagged "Technical_error" were removed. Further details on the phenotype preparation can be found on the UK Biobank showcase website (http://biobank/ndph.ox.ac.uk/showcase/label.cgi?id=220).

Genotype data were restricted to 464,708 individuals who clustered genetically within a European population group. After filtering for imputation quality and allele frequency, a total of 11,511,739 variants were retained. NMR metabolite GWAS was performed using Bolt-LMM in which variants were included in the random effects component. Association analyses were adjusted by sex, array and fasting time. Genetic instruments associated with each metabolite ($p<5e^{-08}$) were identified and clumped ($r^2<0.001$) using the TwoSample MR package[36]. Pleiotropy was assessed by looking up the association between the genetic instruments for each candidate metabolite (i.e. the potential 'instruments') with all metabolites.