The Emerging Landscape of Epidemiological Research Based on Biobanks Linked to Electronic Health Records: Existing Resources, Analytic Challenges and Potential Opportunities

Authors: Lauren J Beesley[1], Maxwell Salvatore[1], Lars G. Fritsche[1], Anita Pandit[1], Arvind Rao[2], Chad Brummett[3], Cristen J. Willer[2], Lynda D. Lisabeth[4], Bhramar Mukherjee[1]*

Affiliations:
[1] University of Michigan, Department of Biostatistics
[2] University of Michigan, Department of Computational Medicine and Bioinformatics
[3] University of Michigan, Department of Anesthesiology
[4] University of Michigan, Department of Epidemiology

Key Words: biobanks, electronic health records, Michigan Genomics Initiative

*Corresponding author: bhramar@umich.edu
Word count: ~10,700
References: 252
Display items: 9

Abstract

Biobanks linked to electronic health records provide a rich data resource for health-related research. With the establishment of large-scale infrastructure, the availability and utility of data from biobanks has dramatically increased over time. As more researchers become interested in using biobank data to explore a diverse spectrum of scientific questions, resources guiding the data access, design, and analysis of biobank-based studies will be crucial.

The first aim of this review is to characterize the types of biobanks that are discussed in the recent literature and provide detailed descriptions of specific biobanks including their location, size, data access, data linkages and more. The development and accessibility of large-scale biorepositories provide the opportunity to accelerate agnostic searches, new discoveries, and hypothesis-generating studies of disease-treatment, disease-exposure and disease-gene associations. Rather than spending time and money designing and implementing a single study with pre-defined objectives, researchers can use biobanks' existing data-rich resources to answer scientific questions as quickly as they can analyze them. While the data are becoming increasingly available, additional thought is needed to address issues related to the design of such studies and analysis of these data. In the second aim of this review, we discuss statistical issues related to biobank research in general including study design, sampling strategy, phenotype identification, and missing data. These issues are illustrated using data from the Michigan Genomics Initiative, UK Biobank, and Genes for Good. We summarize the current body of statistical literature aimed at addressing some of these challenges and discuss some of the standing open problems in this area. This work serves to complement and extend recent reviews about biobank-based research and aims to provide a resource catalog with statistical and practical guidance to researchers pursuing biobank-based research.

Abbreviations:

BMI = body mass index
EHR = electronic health record
eMERGE = Electronic Medical Records and Genomics Network
GFG = Genes for Good
GREML = genomic relatedness-matrix restricted maximum likelihood
GWAS = genome-wide association study
ICD = International Classification of Diseases
KPRB = Kaiser Permanente Research Bank
MGI = Michigan Genomics Initiative

NIH = National Institutes of Health
MLMA = mixed linear model association analysis
NHGRI = National Human Genome Research Institute
PheRS = Phenotype Risk Score
PheWAS = phenome-wide association study
SNP = single nucleotide polymorphism
UKB = UK Biobank

Section 1: Introduction

Biobanks linked to electronic health records (EHR) provide a rich data resource for health-related research. Biobanks, loosely defined, are biorepositories that accept, process, store and distribute biospecimen and/or associated data for use in research and clinical care.[1] The rise in the number and size of biobanks across the world in recent years can be explained by improvements in biospecimen analysis and the need for large datasets to address complex diseases and conditions.[1,2] Many types of biobanks exist, including commercial, single medical center, health system-based, and population-based biobanks. Some biobanks are disease- or organ-specific, while others encompass a large breadth of diseases.

The development and accessibility of large-scale biorepositories provide the opportunity to accelerate agnostic ("hypothesis-free") searches, new discoveries, and hypothesis-generating studies of disease-treatment, disease-exposure and disease-gene associations. Rather than spending time and money designing and implementing a single study, researchers can use biobanks' existing data-rich resources to answer scientific questions as quickly as they can analyze them. With the establishment of biobank infrastructure, the availability and utility of data from biobanks has dramatically increased over time, and scientific interest in biobank-based research has grown. Moreover, governments and institutions are investing in the establishment of large-scale biobanks such as the US National Institutes of Health's upcoming *All of Us* biobank[3] and the well-established, multi-institutional UK Biobank (UKB).[4,5] As more researchers become interested in using biobank data to explore a diverse spectrum of scientific questions, resources guiding the data access, design, and analysis of biobank-based studies will be crucial. Comprehensive resources describing the types of data available in major biobanks and comparing their patient populations and research emphases are still limited.

Recent reviews briefly discuss statistical and computational considerations for studies involving genetic data,[6] limitations of traditional study designs and identifying real world phenotypes,[7,8] and EHR-based approaches and database linkages in making pharmacogenetic discoveries.[9] These reviews are limited in their discussion of statistical methods related to biobank and EHR-based research and in their exploration of critical concepts such as study design, sampling, missing data, and other analytic issues related to biobank research. In this paper, we complement and extend recent publications about biobank-based research with the ultimate goal of providing an extensive catalog of resources and some practical guidance to researchers pursuing biobank-based research. In Section 2, we characterize different types of biobanks and provide detailed descriptions of specific biobanks including their geographic location, size, data access and availability, data linkages and more. We also discuss the dominant health-related outcomes studied in biobank research to date. In Section 3, we describe different statistical approaches for genome- and phenome-wide association studies (GWAS/PheWAS), an area of particular interest in biobank research. In Section 4, we discuss general statistical issues related to biobank research including study design, sampling strategy, phenotype identification, and missing data. We illustrate some of these issues using data from three biobanks, the Michigan Genomics Initiative (MGI)[10,11] the UK Biobank (UKB)[4,5], and Genes for Good (GFG). In Section 5, we mention potential opportunities and promising future directions for expanded and improved biobank-based research through a discussion of novel and emerging uses of EHR data and the integration of EHR with external data sources.

Section 2: A Characterization of Major Biobanks

In this section, we describe the types of biobanks that are frequently discussed in the literature and provide detailed descriptions for many specific biobanks. We then discuss recent biobank-based literature and highlight specific topics receiving particular attention. The goal of this section is to provide a high-level overview of the kinds of research being conducted using biobank data and to provide practical resources describing specific biobanks.

Existing Biobanks

Table 1 describes some notable biobanks in terms of their size, location, type, and accessible data. This table extends the biobank descriptions in Wolford et al. (2018) to include additional information about data linkages and cohort characteristics, and it includes information for a broader set of biobanks.[6] Many of the biobanks listed in **Table 1** provide access to data for outside researchers. These biobanks are often connected to EHR and contain genotype information for some of the patients. Some of these biobanks also have linkage to death registries and detailed prescription information. The biobanks in **Table 1** often fall into two general categories: population-based biobanks and medical cancer/health care system-based biobanks.

Population-based biobanks

Population-based biobanks are large-scale biorepositories that aim to recruit subjects reasonably representative of the source population. Population-based biobanks recruit directly from the general population (e.g. China Kadoorie Biobank), and subjects are eligible for enrollment irrespective of their disease status or healthcare utilization. Estonia,[12,13] Denmark,[14] Sweden,[15] Saudi Arabia,[16] China,[17] the Republic of Korea,[18,19] Qatar,[20,21] and Taiwan[22,23] are just some of the countries that have invested in establishing population-based (or reasonably representative) biobanks.

Perhaps the most well-known population-based biobank that has been used for research is the UKB[4] With over 500,000 subjects, it is one of the largest biobanks in the world. All residents aged 40-69 who lived within 25 miles of one of their 22 assessment centers (~9.2 million) were invited to participate.[5] UKB takes advantage of the UK National Health Service to obtain follow-up data (e.g. mortality, cancer registrations, hospital admissions, primary care data, etc.) and actively collect and verify conditions that are typically under-reported (e.g. cognitive function, depression).[5]

Medical Center and Health system-based biobanks

Another common type of biobank is based on a medical center or a particular health care system. In general, health system-based biobanks, such as Partners HealthCare Biobank, contain EHR and genotype data along with survey data. Some, like the large Kaiser Permanente Research Bank (KPRB), have additional linkages with detailed prescription information and feature-specific sub-cohorts (e.g. pregnancy and cancer cohorts in the case of KPRB). A notable health-system based biobank is the Million Veterans Program. With already more than 600,000 enrolled, it is one of the world's largest genomic biobanks, and it recruits participants from the US veteran population, allowing for the investigation of military-related diseases and conditions. Other such biobanks recruit patients from a distributed network of health centers throughout the country. Their sampling strategy many include active recruitment for particular subpopulations; for example, BioBank Japan[24] recruits patients with particular current or past disease status and the upcoming NIH *All of Us*[25] program will feature the active recruitment of underrepresented minorities.

MGI (used in illustrative examples below) is an academic medical center-based biobank that started at the University of Michigan in 2012. It recruits surgical patients over the age of 18 based on opt-in consent (allowing for re-contact for future research purposes), collects and stores blood samples, genotypes DNA samples, collects brief survey data related to pain, and is linked to EHR. This biobank also links patient data to other data sources including their cancer registry, prescription data, insurance claims and national death index and is also undergoing an effort to implement a broad epidemiologic questionnaire designed to be comparable to other biobank survey data, namely the UKB. For some biobanks, select biobank descriptives are publicly available online without application; for example, MGI publishes summary counts for International Classification of Diseases (ICD) codes and some PheWAS results,[11] and DiscovEHR shares frequencies of various genetic variants.[26]

Other types of biobanks

Initially planning to become the first nationwide biobank, deCODE Genetics is now a privately-owned commercial biobank. Launched in 2007 and funded by the National Human Genome Research Institute (NHGRI), the Electronic Medical Records and Genomics (eMERGE) Network combines a network of DNA biorepositories linked with EHR as a resource for genetic analyses. Disease-specific biobanks are also common, and these biobanks may focus on rarer conditions. Two examples are PcBaSe Sweden,[27] a prostate cancer cohort, and the Mayo Clinic Biobank for bipolar disorder.[28] While disease-specific biobanks may be better powered than other biobank types to study certain diseases, they are typically smaller, may not be linked with EHR, and may not have genotype data readily available.

GFG (used in illustrative examples later on) is a subject-initiated biobank that started at the University of Michigan in 2015. It recruits participants over the age of 18 from all 50 US states through an online Facebook application, collecting survey data on health and behavior. As an incentive for continued participation and contribution of a saliva sample for genotyping, participants also receive ancestry analysis and the option to download their raw genetic data. At the time of publication, over 70,000 participants are enrolled and over 27,000 have been genotyped. Unlike many other biobanks in this paper, GFG is not linked to EHR data.

Recent Major Biobank-Based Literature

In order to characterize the current biobank literature, we conducted a brief literature search using PubMed to find papers about biobanks and biobanking and papers using biobank data. Details regarding the methodology used to identify publications can be found in **Supplementary Section S1**. We emphasize that this is not intended to be an exhaustive list of biobank-based literature. The papers published about biobanks or using biobank data can be roughly grouped based on their scientific goals as follows: (1) biobank study design and cohort characteristics, (2) ethics and public perception of biobanks, (3) feasibility and implementation, (4) exploration of treatments and therapies, (5) epidemiologic exploration focused on non-genetic data, and (6) epidemiologic exploration using genetic data. Below, we review papers in these six broad categories in more detail.

Study Design and Cohort Characteristics

Biobanks typically publish papers on study design,[24,29,30] cohort characteristics,[13,30–34] how the cohort differs for the rest of the country's population,[35] and characteristics of specific patient populations (e.g. clinical characteristics of colorectal[36] and prostate[37] cancer patients in the BioBank Japan cohort). This information is critical for determining generalizability of results obtained using biobank data.

Ethics and Public Perception

There has been a good deal of attention given to ethics of biobanks, particularly ethical and legal concerns[2] with the use of broad consent (seeking consent for future unspecified research). Particular attention has been given to the use of opt-out consents in biobanks with plans for broad, long-term use.[38,39] Additionally, research has looked at the public perceptions of biobanks and biobanking,[40] identified areas of reluctance for potential subjects to consent, and gathered general thoughts on medical and epidemiological research. While hurdles exist (including concerns about privacy and confidentiality, benefit-sharing and commercialization, and internationalization), there is evidence from Germany[41] and China[42] that there is general public support for biobanks and large-scale cohort studies.

doi:10.20944/preprints201809.0388.v1

Feasibility and Implementation

Literature about biobanks explores feasibility and implementation for establishing biobanks, including business plans and models for facilitating biobank creation,[43] how to recruit and obtain consent (particularly among particular groups of patients such as cancer patients),[44–46] and the use of electronic consent in biobanking.[47] Increasingly, biobanks are augmenting their survey data with EHR database connections. The promise and utility of EHR data for secondary research use has been well-established.[48,49] Research into EHR data quality suggests a need for standardized methods of EHR data quality assessment[50] and awareness of underlying data collection processes.[51] Concerns around EHR data manipulation and analysis are discussed later.

Scientific Studies of Health-Related Outcomes

The vast majority of emerging biobank-based literature focuses on studying health-related outcomes. One area of exploration involves comparisons or characterizations of different *treatments or therapies*. For example, Ramirez et al. (2012) examined the impact of genetic variants in European-Americans and African-Americans on the response to different warfarin dosages.[52] EHR-linked biobank data is well positioned to explore treatment or therapy outcomes, treatment repurposing, and gene-by-treatment interactions.

Other studies use biobank data to perform epidemiologic analyses using available EHR and/or supplemental survey data.[32,53,62–71,54,72–75,55–61] We group these papers published using biobank data into two coarse categories: *genetic* and *non-genetic analyses*. Two examples of non-genetic analyses include Song et al. (2018), which describes the protective nature of alcohol consumption on coronary artery disease risk in the Million Veterans Program, and Peters et al. (2018), which describes sex differences in the association between measures of general and central adiposity and risk of myocardial infarction in the UKB.[54,55] Pilling et al. (2017) is an example of a genetic study, where the authors conducted a genome-wide association study of UKB data to identify 25 loci associated with human longevity.[76] Another example of a genetic analysis paper, Nielsen et al. (2018) used biobank data to explore the relationship between genetics and atrial fibrillation.[77]

Figure 1 provides a distribution of included biobank-based publications falling into each of the above categories over time. The recent, rapid increase of biobank-based analyses, particularly the non-genetic and genetic-based analyses of health-related outcomes, is evident. The rise of genetic-based studies can be partly explained by the increase in the number of genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS). GWAS use genotype data, typically from a large number of individuals, to relate millions of genetic variants with a given disease/health condition, and biobanks often contain upwards of several hundred thousand individuals. Additionally, many biobanks have linked the genotype data to EHR, which allows for in-depth phenotyping and, thus, the feasibility of relating millions of genetic variants with hundreds of diagnoses and lab tests, leading to exploration of the genome x phenome landscape through PheWAS.

While the overall number of biobank-related papers has been increasing rapidly, it is worth exploring the number and types of publications produced by individual biobanks. The types of papers published for a particular biobank may depend on the kinds of data available and the willingness to share data externally. **Table S1** provides additional details about the types of identified papers associated with several prominent biobanks. UKB is associated with a large number of publications and particularly papers involving genetic data. The large volume of publications can be explained by external data accessibility and the presence of high-quality genetic information on a large number of patients. In studies conducted using data from other biobanks, UKB data is often chosen as a validation dataset.

Common Outcomes in Biobank Research

While data in large biobanks allow researchers to examine a broad array of outcomes (and often many at once), psychiatric/neurologic outcomes, cardiovascular disease, obesity/diabetes, cancer, and pulmonary conditions dominate recent biobank-based research. Common psychiatric and neurologic outcomes include risk-taking behavior,[78,79] depression/major depressive disorder,[78,80,89,81–88] Alzheimer's disease,[22,81,87,88,90] anxiety,[78,79,82] schizophrenia,[78,82,83,87,88] and bipolar disorder.[78,82,83,87,88,91–93] These outcomes are ascertained by either diagnosis codes or survey responses, and different definitions and thresholds are used in sensitivity analyses. Similarly, cardiovascular disease outcomes include coronary artery disease/coronary heart disease, [32,54,98–100,55,64,87,88,94–97] which are defined as a combination of more specific conditions including myocardial infarction. Related conditions like stroke, atrial fibrillation[77,101–103] and calcific aortic valve stenosis[104] have also been explored in the literature. Obesity (and related measurements like BMI and waist-to-hip ratio) and diabetes have also been explored.[54,58,110–118,95,99,100,105–109] Colorectal,[53,119] breast,[57,120,121] lung,[72] pancreatic,[122] and skin[10] cancers as well as pulmonary conditions including smoking[32,59,60,75,123] and airflow obstruction[59,60,124] have been investigated, but to a lesser extent.

While psychiatric/neurologic conditions, cardiovascular disease, obesity, cancer, and pulmonary conditions are responsible for a significant portion of morbidity and mortality, the breadth and depth of EHR-linked biobank data offer a valuable resource to research many other rare and chronic diseases and conditions as well as risk factors and health behaviors. As such, there is great opportunity for future explorations into health outcomes using EHR-linked biobank data.

Section 3: Brief Summary of Statistical Approaches for GWAS/PheWAS

The combination of large-scale genotype and phenotype data provides new avenues for exploring scientific questions regarding the relationships between genotypes and phenotypes. First demonstrated in Denny et al. (2010), PheWAS explore the associations between a single variable of interest and many EHR-derived phenotypes.[125] PheWAS usually relate phenotypes to a single genetic variant or a polygenic risk score (e.g. Fritsche et al. 2018)[10], but it is worth noting that PheWAS can be conducted based on additional biomarker values/lab tests (e.g. Liao et al. 2017, Neuraz et al. 2013).[126,127] This provides a broad range of scientific questions that can be explored through GWAS and PheWAS-type analyses using biobank data. For both GWAS and PheWAS, phenotypes are often defined using ICD codes derived from the EHR (see the section on "Defining the Phenome" for more details).

The current GWAS/PheWAS literature features studies that fall into multiple different categories in terms of their analytic goals. A natural and common goal of GWAS and PheWAS is to study the associations between specific phenotypes and variants at a particular gene region. This analysis is often performed using linear or logistic regression (recently, Firth-corrected logistic regression) or using mixed linear model association (MLMA) analysis.[10,86,95,101,117] A discussion of MLMA and related issues can be found in Yang et al. (2014).[128] Recently, Dey et al. (2017) proposed a fast alternative to Firth-penalized regression to stabilize estimation for PheWAS studies using saddle-point approximation that is particularly useful for handling extremely unbalanced case-control data.[129] Recently, a saddle-point approximation approach for estimating mixed models (called SAIGE) was proposed for handling highly unbalanced case-control data with additional sample relatedness, which is typical for biobanks.[130]

In the PheWAS setting, researchers may be further interested in studying the association between multiple phenotypes in terms of underlying genetic risk. In Fritsche et al. (2018), researchers approach this task by developing a polygenic risk score for a primary phenotype of interest and relating the polygenic risk score to other phenotypes.[10] Another common strategy for identifying phenotype relationships through shared genetic risk is bivariate linkage disequilibrium regression.[86,88,131–133]

Another common target for these studies is to identify the proportion of variation in a particular phenotype that can be attributed to genetic variation, called heritability. Some popular statistical methods include polygenic profile scoring, univariate linkage disequilibrium regression, and genomic relatedness-matrix restricted maximum likelihood (GREML).[86,88,131,134–136] Recently, Bastarache et al. (2018) developed a phenotype risk score-based method (called PheRS) to study rare genetic variants associated with Mendelian diseases.[137]

Recently, researchers have used particular genetic variants as instrumental variables in PheWAS analyses, where loci related to a primary phenotype are selected and their associations with secondary

doi:10.20944/preprints201809.0388.v1

phenotypes are evaluated.[101] Mendelian randomization analysis is then used to explore potential causal relationships between the genetic trait and the primary and secondary phenotypes.[138]

Section 4: Statistical Issues Related to Biobank Research

Study Design

A key issue to consider when performing a biobank-based study is study design. Design choices can have implications for the analysis and interpretation of the study results. In this section, we describe several common approaches for study design used in biobank research and describe some analytical and design-based strategies for dealing with common sources of bias. A critical part of study design is the mechanism by which patients are sampled from the population of interest. Two sampling mechanisms are at play: (1) the mechanism by which subjects are sampled from the population into the biobank and (2) the mechanism by which biobank subjects are sampled for inclusion in the study.

Sampling from the Population

Population-based biobanks like UKB and China Kadoorie Biobank sample patients from a network of health centers or administrative sites across each country. Compared to other types of biobanks, population-based biobanks are often thought to be more representative of the target population and often recruit a larger number of subjects. However, individual characteristics may still impact inclusion in a population-based biobank--for example, living near an assessment center (UKB) or living in a region with certain desirable characteristics (Kadoorie). Medical center-based biobanks (e.g. MGI, BioVU) and health system biobanks (e.g. KPRB, Partners) attempt to recruit all patients that meet certain criteria within the center/health system, often through selected clinics. Generally, participation in these biobanks requires subjects to use healthcare, which is indicative both of ability to access healthcare (e.g. barriers to access including transportation and insurance) and health (i.e. people with diseases and chronic conditions are more frequent users of healthcare). In the case of BioBank Japan, patients at participating health centers are identified if they have had or become diagnosed with one of 47 diseases. Compared to population-based biobanks, academic medical cancer-based biobanks tend to see more patients with rare or complicated diseases due to availability of specialized care and, thus, are often useful for investigating rare conditions. For example, MGI[10,11] is enriched for analyses of some cancer types, most notably melanoma of the skin, since Michigan Medicine is known for its skin cancer treatment and care. Disease-specific biobanks are used to examine specific conditions, and in some cases, disease-specific biobanks will also recruit controls (e.g. PcBaSe Sweden). Of note are biobank recruitment methods that recruit self-selected volunteers directly from the general population such as GFG (subject-initiated via Facebook). Many biobanks will further screen interested volunteers in addition to the sampling mechanisms described above (as is planned in the upcoming NIH *All of Us* biobank). In all cases, the study designs have the potential to induce sampling and participation biases into the analysis. This can have implications on the generalizability of the results and impact measures of association.

To demonstrate an impact of different sampling mechanisms, we consider prevalence rates for different phenotypes in three biobanks, MGI, UKB, and GFG. As mentioned previously, MGI is a biobank of over 60,000 patients treated at an academic medical cancer. Patients in MGI were recruited through the anaesthesiology department as patients were preparing to have a surgical procedure. The UKB is a population-based collection of over 500,000 patients. GFG is a self-initiated biobank recruiting subjects via Facebook. MGI and UKB are linked to EHR, while GFG obtains phenotype information via survey self-reporting. All three biobanks are described in **Table 1**, and **Table S2** in **Supplementary Section S4** provides comparisons of the patients in MGI, UKB, and GFG in terms of demographics. The three biobanks have very different sampling mechanisms into the biobank, and we expect the phenotype prevalences in MGI and GFG patients to be quite different both from the general US population and the subjects in the UKB. Phenotypes were defined for MGI and UKB using aggregated versions of ICD codes, called PheWAS codes or phecodes.[139] Phenotypes in GFG were defined based on survey responses. A description of the phenotype generation process can be found in **Supplementary Section S2**.

Table 2 presents prevalences of many commonly-studied diseases in MGI, UKB, and GFG along with published prevalences for their corresponding target populations. MGI often captures subjects with many conditions at a higher rate than is observed in the nationwide population. The UKB has higher case counts for several conditions due to its size, and it is often more representative of the rates observed in the population (at least for conditions common among ages 40-69, the age range included in UKB). We note that there are several diseases for which the ICD-derived phenotype classifications in UKB *do not* appear representative, particularly obesity. We discuss this issue in more detail in **Supplementary Section S5**. The high prevalence of depression in GFG is a result of the broad definition of the depression phenotype, which is obtained by asking subjects whether they have every felt depressed. We note the small proportion of subjects diagnosed with breast cancer and prostate cancer in GFG compared to the US population. This may be a result of the differing age distributions, where GFG consists of generally younger people. Differences between GFG and the other biobanks may be a result of different sampling mechanisms or differing phenotyping procedures.

The difference in sampling mechanisms between the MGI and UKB has an impact on observed disease prevalences for many types of diseases. **Figure 2** shows the relative prevalence of various phenotype codes within different disease categories between MGI and UKB. We see that the majority of the prevalences are higher in MGI. In particular, prevalences for neoplasms, symptoms, endocrine/metabolic disorders, infectious diseases, and congenital anomalies are uniformly higher for MGI compared to UKB.

The biobank sampling mechanism may also have implications for the use of EHR data. Population-based biobanks may be more likely to have access to a patient's *primary care center* EHR but might have to deal with heterogeneity both in terms of the EHR-interface used to collect and store the data and differences in case/procedure/diagnosis reporting.[51,140] Some population-based biobanks may overcome/mitigate many of these issues if they operate in countries with universal healthcare (BioBank Japan) or publicly funded healthcare (UKB). Medical center-based biobanks may face complications related to patients coming to their centers for specialized treatment; for example, cancer surgery. While EHR data related to the observed surgery and treatment would often be robust, we might expect the length of each patient's medical history may be shorter and less complete compared to population-based biobanks, since many patients may return to their local health care provider for post-surgery treatment. Unlike biobanks in an academic medical center, we believe broad health system-based biobanks and population-based biobanks may likely have more detailed and consistent data on biobank subjects and may have more complete EHR data for subjects with more common and easily-managed diseases.

Sampling from the Biobank

Within pre-existing biobanks, researchers then seek to sample patients for inclusion in a particular study. Such samples may be limited by data availability, where some subjects may not have, for example, genotype information or survey response information. A common study design involves phenotype-specific case-control sampling, where all observed cases for a particular phenotype are selected and some subset of (possibly matched) controls for that phenotype are sampled from the biobank (e.g. Fritsche et al. 2018, Abana et al. 2017).[10,120] An advantage of case-control sampling is that it does not require longitudinal information and instead relies on dichotomized phenotypes, but it is heavily dependent on the "case" and "control" definitions.

Another common study design is cohort sampling, where all biobank subjects with available data are included in all analyses (e.g. Au Yeung et al. 2014, Hall et al. 2018).[86,141] Self-controlled designs in which each subject serves as his/her own control are emerging as an appealing design paradigm for some scientific problems (e.g. Kuhnert et al. 2011, et al. Zhou 2018).[142,143] Two variations of self-controlled designs are the self-controlled case series design and the cross-over design. Recently, Schuemie et al. (2016) developed an adapted self-controlled case series design that uses the notion of accumulated exposure to study long-term drug effects.[144] A detailed comparison of the two primary designs can be found in MacClure et al. (2012),[145] and additional exploration of self-controlled case series can be found in Petersen et al. (2016) and Simpson et al. (2013).[146,147] An advantage of this design is that it controls for confounding due to time-invariant variables. Unlike cohort and case-control designs, however, this method requires longitudinal data to be available for all subjects. Large-scale longitudinal observational databases, such as EHR-linked biobank databases with time-stamped diagnosis, procedure and therapy data, are readily accessible resources for many longitudinal outcomes.[146] However, self-controlled designs require adequate longitudinal data (in terms of number of visits and length of follow-up), which can either be missing or incomplete in some EHR-linked databases.

Impact of Sampling Mechanism on Inference

Madigan et al. (2013) compares effect estimates resulting from self-controlled case series, cohort, and case-control designs in a particular setting and demonstrates that the choice of study design can have substantial impacts on effect estimates.[148] These choices also impact the statistical power and generalizability of the results. Therefore, study design should be considered carefully. In addition to impacting power, the method by which the subjects are chosen may result in biased inference (with respect to the target population), called sampling bias. Haneuse et al. (2016) provides a general framework for exploring and dealing with selection/sampling bias for EHR-based analyses.[149] Haneuse et al. (2016) focuses on characterizing the mechanism by which subjects were included in the dataset by breaking it into smaller observation mechanisms. For example, a subject may be included in a study if 1) the subject is selected for inclusion in the biobank, 2) the subject consents, and 3) the subject is selected from the biobank by study researchers. Different factors may impact different selection mechanisms, and possible sources of selection bias arising from each individual step can be explored in detail in a sensitivity analysis framework.

The impact of the selection procedure on inference may depend on the analysis being performed. For example, case-control sampling from the biobank will result in biased estimates of the marginal probability of having a disease; however, this sampling design may be able to produce valid estimates of the association between disease status and a covariate. Statistical methods exist for addressing some sampling biases using, for example, inverse probability weighting.[150] Some recent works exploring selection/observation biases in the EHR setting include Zheng et al. (2017), Phelan et al. (2017), Goldstein et al. (2016), and Rusanov et al. (2014).[151–154]

There is a belief in the literature that GWAS/PheWAS study results may be less susceptible to bias resulting from the patient sampling mechanism, but bias due to genotype relationships with the sampling mechanism can still arise in certain settings.[155,156] Additional work may help clarify settings in which bias is and is not expected in GWAS and PheWAS studies. In general, issues of sampling bias are not unique to EHR data, and many authors have explored the impact of sampling on inference. However, additional characterizations of the mechanisms by which we can have sampling bias in biobank and EHR research may help guide study design in the future.

Dealing with Confounding

In addition to sampling, measured and unmeasured confounding are common sources of bias in observational data. Careful use of existing analytical tools can help reduce or eliminate biases resulting from confounding. Here, we define a confounder as a variable that impacts both our outcome and our predictor(s). We exclude the situation where the variable is a mediator. Failure to adjust for the confounder may result in biased inference regarding the association between the predictor and the outcome. In a given dataset, sampling and confounding biases can both be present, and careful adjustment of one source of bias does not preclude the possibility of bias from the other source. Haneuse (2016b) details differences between sampling and confounding biases, where sampling bias resulting from the patient selection mechanism impacts external validity of the results, and confounding biases impact internal validity.[157] There are many analytical strategies in the statistical literature for dealing with confounding. A typical method is to adjust

doi:10.20944/preprints201809.0388.v1

for confounders in a statistical model or stratify analyses by the potential confounders (e.g. Hall et al. 2018).[86] Techniques for reducing and eliminating confounding often assume that the potential confounders are measured. In the EHR setting, however, some confounders of interest (e.g. comorbidities) may often be unmeasured, crudely measured, or incomplete. In such settings, sensitivity analyses and related statistical methods can be used to explore the impact of and to correct for potential unmeasured confounding.[158–160]

Biobank data provides several design-based strategies for dealing with confounding as well. In a case-control sampling framework, controls can be matched to cases based on potential confounders such as age and gender, which can make the case and control populations more similar in terms of their age and gender distributions (e.g. Fritsche et al. 2018).[10] Rather than stratifying statistical analysis by a potential confounder, one could directly define the analytical sample within narrow windows of a particular confounder (e.g. subjects ages 60-65). With large biobank datasets, we can often still obtain an analytical sample of a substantial size with narrow inclusion constraints. As mentioned previously, self-controlled studies adjust for time-invariant confounders through the design, and additional statistical methods have been developed to further account for systematic differences between time periods.[161] In terms of methods designed for large-scale agnostic EHR-based studies such as GWAS or PheWAS, Schuemie et al. (2014) and Schuemie et al. (2018) propose a p-value calibration method that may be able to account for both random and systematic (e.g. confounding, sampling biases) sources of error using distributions of effect estimates believed to be truly null effects.[162,163]

Additional Thoughts on Identifying the Study Sample

An additional concept to consider when defining the study sample is the independence between subjects. Longitudinal outcomes are expected to be correlated within patients, and outcomes may be correlated between patients due to relatedness, nesting within doctor or clinic, belonging to a common social network, or other reasons. The software KING (Kinship-based Inference for GWAS) uses genotype data to determine pairwise kinship between subjects.[164] We might then define the study sample restricted to unrelated subjects and apply methods that rely on independence between subjects (e.g. Firth-corrected logistic regression in Fritsche et al. 2018).[10] Statistical modeling approaches such as mixed modeling can also be used to account for residual correlations between individuals.[130]

Although not discussed in detail here, finite resources for collecting patient information presents another sampling-related challenge. In particular, suppose we want to collect genotype information on some subset of our subjects. Who do we test? This and related issues are explored in detail in Sun et al. (2017)[165] and Schildcrout et al. (2015) and (2018).[166,167]

Defining the Phenome

A central challenge for research involving EHRs is in defining phenotypes. The data available falls into two broad categories: structured and unstructured. Some examples of structured data are billing and procedure codes, numeric lab and test results, and prescription information (both what has been prescribed and what has been filled). Some examples of unstructured data are the narrative notes made by physicians/nurses and radiological/pathological notes and images. For a detailed review of phenotyping procedures, see Bush et al. (2016).[8]

Phenotypes from Structured Data

Previous PheWAS studies primarily rely on structured data to define the phenotypes. In particular, ICD9 and ICD10 diagnosis codes (International Classification of Diseases, revisions 9 and 10) are the most common source used for defining phenomes.[168] These codes are appealing due to their standard definitions (although perhaps with differential usage in practice) across institutions. These codes are often aggregated to conform to a standardized set of phenotype definitions, called "phecodes." However, there is a large amount of additional information in the EHR that can be used to define phenotypes. **Figure 3** provides some examples of the types of structured and unstructured EHR information that can be used to construct phenotypes.

There are many challenges to incorporating additional structured EHR information to define the phenotypes. One challenge involves automation and computation. Suppose we define phenotypes based on structured EHR-based measures of many different types, e.g. binary, count, and continuous. For example, suppose we have many continuous lab values. We may be tempted to model all values using linear regressions, but pre-processing may be required to determine whether the linear regression assumptions are reasonably met. Such evaluation may be difficult to perform manually for a large number of phenotypes, and the development and use of automated algorithms is essential.[169] Another issue involves comparability of the phenotypes between institutions, where lab tests may be performed using different assays or with different rates of variability, and there may be differences in coding and billing norms.

An alternative strategy to phenome generation uses additional expert input (for example, through a consortium) to inform the phenotype definitions. However, establishing a well-accepted definition for a given phenotype requires time, careful thought, and discussion. The eMERGE Phenotype Knowledgebase[170] (PheKB) details existing phenotyping algorithms for individual phenotypes that incorporate additional patient information. Due to the complexity of these phenotyping algorithms, the simpler ICD-based phenotyping method is common for PheWAS studies, but incorporation of these external phenotyping resources may help improve phenotype definitions in the future.

Phenotypes from Unstructured Data

Unstructured data has also been used to define phenotypes, particularly for diseases with unreliable ICD9 classifications such as some psychiatric diseases, using natural language processing methods.[171–179] Such methods can also be used to obtain patient measures such as smoking status.[171] Natural language processing methods mine free text such as narrative doctor’s notes for words or phrases corresponding to a particular characteristic. The general goal is to develop a model combining structured and unstructured data to classify each patient as having or not having the phenotype of interest in such a way as to maximize prediction abilities for the sample as a whole, perhaps measured using negative or positive predictive value.[171,172] Some challenges include dealing with misspellings, tenses, alternative phrasing, and defining a trained dictionary of words and phrases that may correspond to a particular phenotype. Algorithms are usually trained using expert annotations, but recent methods have attempted to automate this step as well.[177,178] Additional machine learning methods have also been used to define phenotypes (e.g. imaging analytics from medical imaging datasets) using a broad spectrum of patient information.[180–182]

Generally, there is a great potential for incorporating data of different types in order to define phenotypes used in EHR-based research. However, future work is needed to provide automated methods for incorporating data of different types for phenome generation.

Phenotype Misclassification

Misclassification of ICD Codes

A common strategy when defining disease phenotypes is to list a subject as being a case if he/she has received a certain number of ICD9/ICD10 diagnosis codes (or composites, called phecodes) for a particular disease. This general strategy, however, only captures part of the story. This disease status determination is usually performed across subjects who have different amounts of follow-up time, who have different numbers of visits, and who are being seen in different types of medical clinics. These factors may all be related to the underlying disease status, and a person who would eventually develop the disease or had developed the disease prior to the follow-up window may not be captured.[183] Some statistical tools have been developed to try to deal with this and related issues, but computational restrictions may make these methods difficult to apply to large-scale biobank data (e.g. Bergeron et al. 2018 and Sinnott et al. 2014).[176,184] Additionally, symptoms occurring between visits may not always be reported, and the use of diagnostic guidelines and assessment of the phenotype may vary from doctor to doctor.[185,186] These underlying patient-specific properties are often ignored when classifying subjects as cases and controls for a particular disease, and this can lead to phenotype misclassification. Such misclassification can be viewed in the context of missing data as explored in the next section.

ICD-based phenotype misclassification is particularly common for psychiatric disorders, where diagnosis can be particularly challenging.[174,185] For diseases with burdensome treatments such as cancer, we may expect that all subjects receiving a cancer diagnosis truly do have cancer, and there may be only a few cancer cases without a corresponding ICD9 code. In contrast, ICD9 codes for psychiatric disorders such as anxiety may be sometimes attributed to some subjects that do not meet the ICD9 definitions for the disorder. There may also be a tendency for patients to receive ICD classifications that result in re-imbursement from the insurance provider. Some ICD codes, for example obesity, may not result in reimbursement and may be expected to have different patterns of misclassification. Additionally, disease ICD codes are sometimes assigned when a disease is suspected prior to further diagnostic testing, so it may be unclear whether a given ICD code refers to the final diagnosis.[8,103]

Phenotype misclassification can result in bias ("information bias") and negatively impact the statistical power to detect associations with the disease of interest. The extent of misclassification can be described using quantities such as sensitivity, specificity, and negative and positive predictive values (provided a gold standard exists for comparison), but these quantities can vary from population to population and from phenotype to phenotype.[187] Therefore, it is difficult to detect the extent of phenotype misclassification in a particular population without performing further phenotype validation.[188] For example, Liao et al. (2017) estimated misclassification rates for particular phenotypes by sampling subsets of patients for manual chart review to verify the phenotype classification.[126] Recently, Huang et al. (2018) explored a method for accounting for phenotype misclassification in association studies using a likelihood-based method that integrates over unknown sensitivity and specificity parameters, placing less emphasis on previously-reported values for sensitivity and specificity from other populations.[188] Duffy et al. (2004) proposes an alternative method for correcting logistic regression effect estimates under misclassification of the outcome.[189,190]

Misclassification in Self-Reported Measures

Another source of phenotype misclassification results from reliance on self-reported measures. For example, self-reported race/ethnicity has been shown to be generally consistent with genetic ancestry but not very specific, particularly for African Americans and Latinos.[191,192] Spangler et al. (2015) reported a discrepancy between self-reported oral contraceptive use with filled prescription data in a population-based study, with prescriptions being filled 11-45% higher than self-reported oral contraceptive use for the same time period.[193] Sensitive health issues may be particularly susceptible to being under- or over-reported, and studies (e.g. those recruiting via social media like GFG) involving such measures should carefully consider the potential impacts of under- or over-reporting on their results.

Missing Data

Missing data is a common issue for biobank analyses, and data may be missing for a variety of reasons. A common source of missingness in GWAS/PheWAS studies is missingness in the genotypes. This is often handled by first excluding subjects with missingness rates above a particular threshold (say, 2%) and then imputing missing values for subjects with smaller missingness rates.[86,88] While many of these biobank analyses reported their treatment of missing genotype data, missing information in the phenotype information or demographics is rarely discussed. For example, when a phenotype is constructed using survey data, how is survey non-response handled? Additionally, many studies define their analytical sample based on some subset of biobank participants. However, it is sometimes unclear how these participants were chosen. A more transparent description of how the study sample was derived and the treatment of missing data may shed some light on the generalizability of study results.

doi:10.20944/preprints201809.0388.v1

Statistical methods for dealing with missing data in the EHR often rely on multiple imputation, a statistical approach in which the missing data is "filled in" using information from subjects with observed values.[194–197] These methods may also rely on natural language processing to obtain information from unstructured clinical notes. Such approaches can prove extremely valuable to EHR-based research, but implicit assumptions about the missingness mechanisms should be carefully considered.

A common type of "missing" data is the true phenotype state of each subject. We can view the sampling mechanism that gave rise to our study population and the mechanism behind phenotype misclassifications (which we might call the observation mechanism) in a missing data framework. The observed phenome in our sample is a function of the true phenome state (the "missing" data), the mechanism by which subjects are sampled, and the mechanism by which phenotypes are observed in the sample as shown in **Figure 4**, where arrows represent dependence.

The probability that a particular subject has an observed phenotype will be related to whether the subject truly has the phenotype, but it may also be related to other factors such as the number of visits to the health care provider, the length of follow-up, the types of health services they receive, and other predictors. These other factors may also be correlated with the true disease status of the subject. For example, a healthier subject may "drop out" of the biobank and may instead seek health care from a tertiary care center. **Figures S3-S5** present descriptions of the length of follow-up, number of unique observed phecodes, and number of visits by gender and observed cancer status in MGI. These figures demonstrate a relationship between these variables and whether the subject ever received an ICD code for cancer during follow-up. The sampling and observation mechanisms and their relationships to underlying disease status and patient characteristics may impact study inference. Further work should be done to explore the impact of different sampling and phenotyping mechanisms on statistical inference.

Multiple Testing

GWAS/PheWAS studies and many other types of EHR-based research often involve the simultaneous testing of many hypotheses. Failure to account for multiple testing can result in inflated Type I error, and many statistical methods have been developed to control the Type I error in the multiple testing setting. Some commonly-used examples include Bonferroni adjustment, false discovery rate-controlling thresholds (e.g. Li et al. 2018),[101] and Benjamini-Hochberg thresholds (e.g. Liao et al. 2017).[126] However, many of these methods (in particular, the simple Bonferroni adjustment method) have been shown to be overly conservative when the many statistical tests are not independent. This is often the case in large-scale GWAS/PheWAS studies, where associations are explored between many different combinations of related characteristics. In this setting, the goal may be to control for the effective number of independent tests rather than the number of correlated tests being performed. Such an approach may improve statistical power to detect significant associations while still controlling the Type I error rate.

Several methods have been proposed to estimate the effective number of tests (e.g. Li 2012) or control for correlated tests. Good (2005) describes resampling-based testing via permutation or bootstrap to correct the p-values for multiple testing.[198] Gao et al. (2008) proposes the simple *M* method to estimate the effective number of tests, which uses a combination of principal components analysis and Bonferroni correction.[199] For a PheWAS study presented in Ge et al. (2017), the effective number of tests is estimated using principal components analysis of a matrix of pairwise correlations between pairs of phenotypes.[135] Alternative methods adjust for multiple testing using multivariate normal assumptions for the correlated test statistics (e.g. Han et al. 2009, Lin 2005, Seaman et al. 2005).[200–202] In the context of correlated SNPs, some methods correct for multiple testing via analysis of the underlying linkage disequilibrium structure of the genetic data (e.g. Duggal et al. 2008).[203] Johnson et al. (2010), Zhang et al. (2012) and Li et al. (2012) provide some simulations comparing the performance of different methods.[204–206]

doi:10.20944/preprints201809.0388.v1

Heterogeneity between Biobanks

Researchers often attempt to validate statistical findings from their data analysis using an independent dataset, often from a different population. Differences between the population characteristics, however, could impact the generalizability of results between populations and impact our ability to replicate findings. Additional issues can arise when comparing inference across datasets with different sampling mechanisms. In the meta-analysis literature, heterogeneity between studies is broadly grouped into three categories: clinical heterogeneity (differences in patients, interventions, and effects), methodological heterogeneity (differences in study design and sampling), and statistical heterogeneity (when the observed effects are more variable across studies than we would expect from random chance). Statistical heterogeneity may be a result of clinical and/or methodological heterogeneity.

Methodological Heterogeneity

To demonstrate the impact of different sampling mechanisms across biobanks on statistical inference (an example of methodological heterogeneity), we consider phenotype co-occurrence rates and genotype-phenotype associations in MGI and UKB. Suppose we are interested in comparing the odds ratio for having a particular phenotype based on the status of another phenotype, called phenotype co-occurrences, in MGI and UKB. While prevalences will clearly be impacted by the different sampling designs between MGI and UKB (see **Figure 2**), it is not clear how the resulting phenotype associations will compare between datasets.

In **Figure 5**, we show the estimated log-odds ratios of having a phecode diagnosis of breast cancer based on other diagnoses in the phenome. See **Supplementary Section S2** for more details on the phenotype generation procedure. We note that the estimated odds ratios from the UKB data tended to be larger in magnitude compared to the odds ratios in MGI. One possible explanation for this phenomenon is that in order for subjects to get a phecode in UKB, they must visit a health care provider, during which time they may get multiple codes. When we compare these subjects with UKB subjects who did not visit a health care provider or did not visit as often, we may obtain inflated odds ratios. The subjects in MGI are enriched with phecodes across the board, but subjects with and without a particular phenotype may have many opportunities to collect other diagnoses through their interactions with the health care provider. In this breast cancer example, the odds ratios for other neoplasms and genitourinary diseases did not exhibit the same differences in MGI and UKB as with other classes of diseases. This may be due to enhanced screening of these diseases after diagnosis of breast cancer in both MGI and UKB. Similar exploration for melanoma showed odds ratio inflation in UKB compared to MGI for all disease categories except neoplasms and dermatological conditions. The odds ratio inflation phenomenon seen for breast cancer and melanoma was present for chronic conditions such as hypothyroidism as well as emergent conditions such as concussions. While the size of the odds ratio estimates differed between the two biobanks, we note that, when both associations were significant at a p-value threshold of 0.05, the associations were largely in the same direction.

There are some associations that may not be appreciably impacted by the sampling mechanism. For example, suppose we are interested in studying associations between various SNPs and a phenotype in a GWAS study. It may be reasonable to believe that any given SNP alone is not appreciably related to selection into the sample or variables related to selection into the sample. In this case, we may believe that GWAS results will be reasonably representative of the population. In **Figure 6**, we compare GWAS results in MGI and UKB for several cancers. In this figure, points represent SNPs identified as being related to the corresponding phenotype in the NHGRI-EBI GWAS catalog. See **Supplementary Section S3** for details. While MGI and UKB have very different sampling mechanisms, the GWAS results generally appear similar between MGI and UKB.

However, this may not always be the case. In **Figure 7**, we compare GWAS results in MGI with results in GFG. We also compare results in both biobanks to genotype-phenotype associations reported in the NHGRI-EBI GWAS catalog. We note that MGI has nearly double the sample size of GFG, and we do not account for this when comparing effect estimates (N = 30,702 vs. 15,156). We do see some differences when comparing GWAS results between MGI and GFG. There are many possible explanations for this. One explanation is that the phenotypes are defined differently in MGI and GFG. In MGI, phenotypes are derived based on ICD codes reported in the EHR during patient follow-up. In GFG, the breast cancer phenotype is derived from survey responses, which ask subjects whether they have ever had breast cancer. The difference in the sampling mechanism both in terms of obtaining subjects and in terms of self-reporting of phenotypes in GFG along with the low number of events in GFG could also explain differences in the GWAS results. A comparison of the MGI GWAS results with the log-odds ratios reported in the NHGRI-EBI GWAS catalog shows a positive relationship, and this relationship is weaker when comparing GFG results with the GWAS catalog estimates.

Clinical Heterogeneity

In addition to differences in the sampling mechanism, differences in patient populations in terms of potential effect modifiers (e.g. age and race) could impact replicability of results across biobanks. For example, suppose we are interested in a particular genotype-phenotype association but that the association varies across genetic ancestries. This is an example of clinical heterogeneity. Such a difference in association could be driven by true biological heterogeneity or by different linkage disequilibrium properties between the populations. When comparing this association overall between two different populations, a failure to adjust for the genetic ancestry composition of the two populations could result in biased inference. Au Yeung et al. (2014) explores the association between ALDH2 and lung function in a southern Chinese population.[141] The authors discuss lack of consistency between their results and results from Western populations, which could be the result of different health attributes of the populations (e.g. different alcohol and smoking behaviors) and could be attributed to different rates of polymorphism between the two populations. An example of this for MGI and UKB is age, where MGI consists of patients aged 18 and up, while UKB consists of subjects aged 40-69. If the association of interest depends on age, we would have different marginal associations in MGI and UKB.

Statistical Methods for Dealing with Heterogeneity

In the presence of this heterogeneity between study populations, we may explore statistical methods to improve our ability to compare between different populations. There is a body of statistical literature for quantifying and handling between-study heterogeneity for meta-analyses (see Thompson et al. 1994, Fletcher et al. 2007, Higgins et al. 2003, and Kriston et al. 2013 for more information).[207–210] Heterogeneity is often handled in meta-analyses through mixed effects modeling. Weighting-based and resampling-based methods for dealing with heterogeneity have also been explored.[211–213] Future work may explore resampling-based methods to make studies more comparable in the presence of heterogeneity with respect to the sampling mechanism.

Section 5: Emerging Uses of Electronic Health Record Data and Combination with External Data

Many of the existing large biobanks in the US are from academic institutions, which may only provide specialty care. Therefore, the EHR from single institutions or health systems may lack the data for some longitudinal analyses. There is a large opportunity to incorporate additional data sources or types to enrich the typical EHR data and enhance the scope of biobank research. For example, by linking cancer and death registry information to the EHR, we may be able to study survival and disease-related outcomes after clinical diagnoses. Local and national surgical registries offer opportunities for more granular health-related outcomes. When registry data is not available, claims data may also provide some insight for survival and disease-related outcomes-based research.[214] Recent work has developed methods for defining the exposome based on clinical narrative information in EHRs or based on additional subject-level measurements.[215,216]

Geo-coded data can be used of explore a wealth of exposure information including social determinants of health, neighborhood characteristics, socioeconomic status, and pollution information.[217–222]

In addition to geo-coded and registry data, longitudinal data within the EHR and beyond offers many opportunities for research. The rise of mobile fitness tracking devices also provides an opportunity to incorporate longitudinal health metrics or even use text messages or game performance to define phenotypes.[223,224] Noren et al. (2010) and Noren et al. (2013) use longitudinal health record data to discover temporal patterns, and Boland et al. (2015) considers seasonal/calendar effects related to disease.[161,225,226] Longitudinal EHR data has proven to be extremely useful in the fields of pharmacovigilance, pharmacoepidemiology, and pharmacogenomics.[52,161,227–230] Additional work leverages large-scale medical data to study potential new indications for existing drugs, called drug repurposing or repositioning.[231] Longitudinal EHR data can also be used to develop dynamic predictions for patient prognosis, adverse events, etc. over time.[232–235] Machine learning methods have great potential for prediction based on EHR data.[236]

When combining data from multiple disparate sources, several problems arise. Most notably are issues regarding patient privacy. Additionally, we must consider issues of data processing, rules for linking records for a single subject, etc. Many statistical methods have been developed for linking records corresponding to individual subjects across data sources, and many of these methods explicitly address issues of privacy.[237–241] Statistical methods have also been developed for combining data across distributed data sources where data from individual subjects is not accessible, called distributed regression analysis. These methods involve sharing sufficient statistics of the data (functions of the individual-level data) from which the individual-level data are not recoverable.[242,243] Yang et al. (2013) developed methods for performing meta-analysis based on sufficient statistics from existing GWAS, and similar methods should be developed for PheWAS studies in the future.[244]

Large biobank datasets also provide an opportunity to study different treatment pathways observed for different patients and their corresponding outcomes.[245] Additional components such as treatment nonresponse and treatment adherence can also be explored.[173,246] While studies of treatment response and adherence are certainly not new, the wealth of information provided through EHRs provides opportunities to study treatment-related outcomes at scale. Additionally, these data sources provide a clearer look at treatment-related outcomes *in practice*, which may not always align with treatment-related outcomes under more ideal settings of a clinical trial. Similarly, these data can be used to analyze and/or predict various outcomes to treatments, medications, and/or dosages for different diseases (sometimes stratified by patient characteristics – e.g. race). For example, Delaney et al. (2012) demonstrated clopidogrel resistance for genetic variants in *ABCB1* and *CYP2C19* using EHR-linked data from cardiac patients.[52,247] Similar analyses can be used for drug repurposing as well.

Randomized controlled trials are often considered a gold standard for statistical inference. Researchers have explored approaches for obtaining results more similar to a randomized trial using observational data and, in particular, EHR data. These methods include carefully-defined inclusion/exclusion criteria, use of weighted analyses and propensity score-based approaches, and definition of exposures and outcomes to mimic a trial.[248–251] An exploration of the use of observational data instead of clinical trials for inference can be found in Franklin et al. (2017).[252]

Section 6: Conclusion

Biobanks linked to electronic health records (EHR) provide a rich data resource for health-related research, and scientific interest in biobank-based research has grown dramatically in recent years. As more researchers become interested in using biobank data to explore a diverse spectrum of scientific questions, resources guiding the data access, design, and analysis of biobank-based studies will be crucial. This work serves to complement and extend recent publications about biobank-based research (e.g. Wolford et al. 2018, Glicksberg et al. 2018, Bush et al. 2016, Ohno-Machado et al. 2018) and aims to provide some statistical and practical guidance to researchers pursuing biobank-based research.[6–9]

In this paper, we provide a detailed characterization of many of the major EHR-linked biobanks in an effort to facilitate researchers' ability to obtain and investigate research-quality biobank data with some understanding of the associated population, sampling mechanism, and data linkages. We also survey biobank-based papers that have been published. Papers using biobank data have focused on illnesses and conditions that cause a large portion of morbidity and mortality including cancer, cardiovascular disease, and obesity/diabetes. Future research can utilize increasingly large EHR-linked biobank cohorts to study a broader range of diseases. Biobank data also present an exciting opportunity to explore treatment and therapy schedules, drug repurposing, or gene-by-treatment interactions in the future. Such explorations can also be used to inform dynamic, patient-centric predictions for monitoring and treating future patients.

When using biobank data for health-related research, it is important that researchers understand the statistical and practical issues that accompany such analyses and have resources to address them. We describe many of the statistical challenges involved in biobank research and some current statistical methods. However, there is a great need for further statistical developments to address the many varied issues that go hand in hand with EHR-based research. One large challenge involves defining the phenome. Many methods have been developed to incorporate unstructured EHR data through natural language processing methods or image analytics, and some researchers have considered other issues of misclassification related to ICD9/10-based phenotype classification. Future work can expand on these methods and explore ways to incorporate a broader spectrum of EHR information into phenotype classification.

Missing data is another broad issue with EHR data. Data can be missing for a variety of reasons, and the mechanism generating the missingness can have large implications on inference. Statistical methods tailored to handling issues of missing data in EHR could prove extremely useful. Additional work regarding sampling mechanisms (e.g. into the biobank, into the study, consenting) is needed to clarify in which settings these sampling mechanisms will impact inference.

With an increase in the volume and variety of data becoming available, additional emphasis should be placed on methods for incorporating data from external sources and emerging data streams (for example, geo-coded data, longitudinal biomonitoring data, mobile data, registry data, genomics/metabolomics data, imaging data, ecologic data, etc.). Such analyses can widen the scope of scientific questions we can address, and they necessitate a new wave of related statistical methods.

Acknowledgements

Part of this research has been conducted using the UK Biobank Resource under application number 24460. The research is supported by NCI P30 CA046592.

Contributions

Lauren J Beesley wrote the manuscript, gathered papers regarding statistical methods and issues related to handling and analyzing biobank data, and prepared the figures.
Maxwell Salvatore wrote the manuscript, gathered papers published about biobanks or those using biobank data, and prepared the tables.
Lars Fritsche performed GWAS analyses and detailed paper edits.
Anita Pandit performed GWAS analyses.
Bhramar Mukherjee provided key guidance in the development of the manuscript throughout the entire process and performed detailed paper edits.
Anita Pandit, Arvind Rao, Chad Brummett, Cristen J. Willer, and *Lynda D. Lisabeth* helped edit the paper. All authors reviewed the manuscript.

References

1. De Souza, Y. G. & Greenspan, J. S. Biobanking past, present and future. *AIDS* **27,** 303–312 (2013).
2. Greely, H. T. The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks. *Annu. Rev. Genomics Hum. Genet.* **8,** 343–364 (2007).
3. All of Us Website. Available at: https://allofus.nih.gov.
4. UK Biobank Website. Available at: http://www.ukbiobank.ac.uk.
5. Allen, N. *et al.* UK Biobank: Current status and what it means for epidemiology. *Heal. Policy Technol.* **1,** 123–126 (2012).
6. Wolford, B. N., Willer, C. J. & Surakka, I. Electronic health records: The next wave of complex disease genetics. *Hum. Mol. Genet.* **27,** R14–R21 (2018).
7. Glicksberg, B. S., Johnson, K. W. & Dudley, J. T. The next generation of precision medicine: Observational studies, electronic health records, biobanks and continuous monitoring. *Hum. Mol. Genet.* **27,** R56–R62 (2018).
8. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17,** 129–145 (2016).
9. Ohno-Machado, L., Kim, J., Gabriel, R. A., Kuo, G. M. & Hogarth, M. A. Genomics and electronic health record systems. *Hum. Mol. Genet.* **27,** R48–R55 (2018).
10. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* 205021 (2018). doi:10.1016/j.ajhg.2018.04.001
11. Michigan Genomics Initiative Website. Available at: https://www.michigangenomics.org.
12. Estonian Genome Center. Available at: https://www.geenivaramu.ee/en/access-biobank.
13. Leitsalu, L. *et al.* Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *Int. J. Epidemiol.* **44,** 1137–1147 (2015).
14. Danish National Biobank. Available at: http://www.biobankdenmark.dk.
15. Biobank Sweden. Available at: http://biobanksverige.se/research/.
16. Saudi Biobank. Available at: http://kaimrc.med.sa.
17. China National GeneBank. Available at: https://www.cngb.org/home.html.
18. National Biobank of Korea. Available at: http://www.nih.go.kr/NIH/cms/content/eng/14/65714_view.html.
19. Cho, S. Y. *et al.* Opening of the National Biobank of Korea as the Infrastructure of Future Biomedical Science in Korea. *Osong Public Heal. Res. Perspect.* **3,** 177–184 (2012).
20. Qatar Biobank. Available at: https://www.qatarbiobank.org.qa.
21. Al Kuwari, H. *et al.* The Qatar Biobank: background and methods. *BMC Public Health* **15,** 1208 (2015).
22. Lin, E. *et al.* Association and interaction effects of Alzheimer's disease-associated genes and lifestyle on cognitive aging in older adults in a Taiwanese population. *Oncotarget* **8,** 24077–24087 (2017).
23. Taiwan Biobank. Available at: https://www.twbiobank.org.tw/new_web_en/index.php.
24. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27,** S2–S8 (2017).
25. National Institutes of Health. *The All of Us Research Program*. (2018).
26. DiscovEHR. Available at: http://www.discovehrshare.com.
27. PcBaSe Sweden Website. Available at: http://www.surgsci.umu.se/english/sections/urology-and-andrology/research/pcbase/?languageId=1.
28. Mayo Clinic Biobank for Bipolar Disorder Website. Available at: https://www.mayo.edu/research/centers-programs/bipolar-disorder-biobank/overview.
29. Al Kuwari, H. *et al.* The Qatar Biobank: Background and methods Chronic Disease epidemiology. *BMC Public Health* **15,** 1–9 (2015).
30. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: Survey methods, baseline

characteristics and long-term follow-up. *Int. J. Epidemiol.* **40,** 1652–1666 (2011).

31. Krokstad, S. *et al.* Cohort Profile: The HUNT Study, Norway. *Int. J. Epidemiol.* **42,** 968–977 (2013).
32. Jiang, C. Q. *et al.* Smoking cessation and carotid atherosclerosis: the Guangzhou Biobank Cohort Study--CVD. *J. Epidemiol. Community Heal.* **64,** 1004–1009 (2010).
33. Awadalla, P. *et al.* Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int. J. Epidemiol.* **42,** 1285–1299 (2013).
34. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44,** 1172–1180 (2015).
35. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *Am. J. Epidemiol.* **186,** 1026–1034 (2017).
36. Tamakoshi, A. *et al.* Characteristics and prognosis of Japanese colorectal cancer patients: The BioBank Japan Project. *J. Epidemiol.* **27,** S36–S42 (2017).
37. Ukawa, S. *et al.* Clinical and histopathological characteristics of patients with prostate cancer in the BioBank Japan project. *J. Epidemiol.* **27,** S65–S70 (2017).
38. Simon, C. M. *et al.* Active choice but not too active: Public perspectives on biobank consent models. *Genet. Med.* **13,** 821–831 (2011).
39. Kaufman, D., Bollinger, J., Dvoskin, R. & Scott, J. Preferences for opt-in and opt-out enrollment and consent models in biobank research: a national survey of Veterans Administration patients. *Genet. Med.* **14,** 787–794 (2012).
40. Ahram, M., Othman, A., Shahrouri, M. & Mustafa, E. Factors influencing public participation in biobanking. *Eur. J. Hum. Genet.* **22,** 445–451 (2014).
41. Starkbaum, J. *et al.* Public Perceptions of Cohort Studies and Biobanks in Germany. *Biopreserv. Biobank.* **12,** 121–130 (2014).
42. Chen, H., Gottweis, H. & Starkbaum, J. Public Perceptions of Biobanks in China: A Focus Group Study. *Biopreserv. Biobank.* **11,** 267–271 (2013).
43. Ciaburri, M., Napolitano, M. & Bravo, E. Business Planning in Biobanking: How to Implement a Tool for Sustainability. *Biopreserv. Biobank.* **15,** 46–56 (2017).
44. Mancini, J. *et al.* Consent for Biobanking: Assessing the Understanding and Views of Cancer Patients. *JNCI J. Natl. Cancer Inst.* **103,** 154–157 (2011).
45. Lee, C. I. *et al.* Patients' willingness to participate in a breast cancer biobank at screening mammogram. *Breast Cancer Res. Treat.* **136,** 899–906 (2012).
46. Pillai, U. *et al.* Factors that May Influence the Willingness of Cancer Patients to Consent for Biobanking. *Biopreserv. Biobank.* **12,** 409–414 (2014).
47. Boutin, N. *et al.* Implementation of Electronic Consent at a Biobank: An Opportunity for Precision Medicine Research. *J. Pers. Med.* **6,** 17 (2016).
48. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12,** 417–428 (2011).
49. Botsis, T., Hartvigsen, G., Chen, F. & Weng, C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.* **2010,** 1–5 (2010).
50. Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Informatics Assoc.* **20,** 144–151 (2013).
51. Richesson, R. L., Horvath, M. M. & Rusincovitch, S. A. Clinical Research Informatics and Electronic Health Record Data. *IMIA Yearb.* **9,** 215–223 (2014).
52. Ramirez, A. H. *et al.* Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record. *Pharmacogenomics* (2012). doi:10.2217/pgs.11.164
53. Morris, J. S., Bradbury, K. E., Cross, A. J., Gunter, M. J. & Murphy, N. Physical activity,

sedentary behaviour and colorectal cancer risk in the UK Biobank. *Br. J. Cancer* **118,** 920–929 (2018).

54. Peters, S. A. E., Bots, S. H. & Woodward, M. Sex Differences in the Association Between Measures of General and Central Adiposity and the Risk of Myocardial Infarction: Results From the UK Biobank. *J. Am. Heart Assoc.* **7,** e008507 (2018).
55. Song, R. J. *et al.* Alcohol Consumption and Risk of Coronary Artery Disease (from the Million Veteran Program). *Am. J. Cardiol.* (2018). doi:10.1016/j.amjcard.2018.01.042
56. Ganna, A. & Ingelsson, E. 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study. *Lancet* **386,** 533–540 (2015).
57. Anderson, J. J. *et al.* Red and processed meat consumption and breast cancer: UK Biobank cohort study and meta-analysis. *Eur. J. Cancer* **90,** 73–82 (2018).
58. Arora, T. *et al.* Self-Reported Long Total Sleep Duration Is Associated With Metabolic Syndrome: The Guangzhou Biobank Cohort Study. *Diabetes Care* **34,** 2317–2319 (2011).
59. Lam, K. H. *et al.* Prior TB, Smoking, and Airflow Obstruction. *Chest* **137,** 593–600 (2010).
60. Amaral, A. F. S., Strachan, D. P., Burney, P. G. J. & Jarvis, D. L. Female Smokers Are at Greater Risk of Airflow Obstruction Than Male Smokers. UK Biobank. *Am. J. Respir. Crit. Care Med.* **195,** 1226–1235 (2017).
61. Yokomichi, H. *et al.* Statin use and all-cause and cancer mortality: BioBank Japan cohort. *J. Epidemiol.* **27,** S84–S91 (2017).
62. Okada, E. *et al.* Demographic and lifestyle factors and survival among patients with esophageal and gastric cancer: The Biobank Japan Project. *J. Epidemiol.* **27,** S29–S35 (2017).
63. Cai, Y. *et al.* Road traffic noise, air pollution and incident cardiovascular disease: A joint analysis of the HUNT, EPIC-Oxford and UK Biobank cohorts. *Environ. Int.* **114,** 191–201 (2018).
64. Wood, A. M. *et al.* Risk thresholds for alcohol consumption: combined analysis of individual-participant data for 599 912 current drinkers in 83 prospective studies. *Lancet* **391,** 1513–1523 (2018).
65. Cohn, E. G., Hamilton, N., Larson, E. L. & Williams, J. K. Self-reported race and ethnicity of US biobank participants compared to the US Census. *J. Community Genet.* **8,** 229–238 (2017).
66. Yaghjyan, L., Rich, S., Mao, L., Mai, V. & Egan, K. M. Interactions of coffee consumption and postmenopausal hormone use in relation to breast cancer risk in UK Biobank. *Cancer Causes Control* **29,** 519–525 (2018).
67. Schooling, C. M. *et al.* Alcohol use and fasting glucose in a developing southern Chinese population: The Guangzhou Biobank Cohort Study. *J. Epidemiol. Community Health* **63,** 121–127 (2009).
68. Teleka, S. *et al.* Risk of bladder cancer by disease severity in relation to metabolic factors and smoking: a prospective pooled cohort study of 800,000 men and women. *Int. J. Cancer* (2018). doi:10.1111/joms.
69. Mc Menamin, Ú. C. *et al.* Hormonal and reproductive factors and risk of upper gastrointestinal cancers in men: A prospective cohort study within the UK Biobank. *Int. J. Cancer* **143,** 831–841 (2018).
70. Kunzmann, A. T. *et al.* Model for Identifying Individuals at Risk for Esophageal Adenocarcinoma. *Clin. Gastroenterol. Hepatol.* **16,** 1229–1236.e4 (2018).
71. Celis-Morales, C. A. *et al.* Associations of grip strength with cardiovascular, respiratory, and cancer outcomes and all cause mortality: prospective cohort study of half a million UK Biobank participants. *BMJ* **361,** k1651 (2018).
72. Hatlen, P., Grønberg, B. H., Langhammer, A., Carlsen, S. M. & Amundsen, T. Prolonged Survival in Patients with Lung Cancer with Diabetes Mellitus. *J. Thorac. Oncol.* **6,** 1810–1817 (2011).
73. Gislefoss, R. E. *et al.* Vitamin D, obesity and leptin in relation to bladder cancer incidence and survival: prospective protocol study. *BMJ Open* **8,** 1–6 (2018).
74. Pang, Y. *et al.* Diabetes, plasma glucose and incidence of fatty liver, cirrhosis and liver cancer: a prospective study of 0.5 million people. *Hepatology* **777,** 1–36 (2017).

75. Bjørngaard, J. H. *et al.* Heavier smoking increases coffee consumption: findings from a Mendelian randomization analysis. *Int. J. Epidemiol.* **46,** 1958–1967 (2017).
76. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany. NY).* **9,** 2504–2520 (2017).
77. Nielsen, J. B. *et al.* Genome-wide Study of Atrial Fibrillation Identifies Seven Risk Loci and Highlights Biological Pathways and Regulatory Elements Involved in Cardiac Development. *Am. J. Hum. Genet.* **102,** 103–115 (2018).
78. Strawbridge, R. J. *et al.* Genome-wide analysis of self-reported risk-taking behaviour and cross-disorder genetic correlations in the UK Biobank cohort. *Transl. Psychiatry* **8,** 39 (2018).
79. Du Rietz, E. *et al.* Association of Polygenic Risk for Attention-Deficit/Hyperactivity Disorder With Co-occurring Traits and Disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 1–9 (2017). doi:10.1016/j.bpsc.2017.11.013
80. Wigmore, E. M. *et al.* Do regional brain volumes and major depressive disorder share genetic architecture? A study of Generation Scotland (n=19 762), UK Biobank (n=24 048) and the English Longitudinal Study of Ageing (n=5766). *Transl. Psychiatry* **7,** e1205 (2017).
81. Gibson, J. *et al.* Assessing the presence of shared genetic architecture between Alzheimer's disease and major depressive disorder using genome-wide association data. *Transl. Psychiatry* **7,** e1094 (2017).
82. Ward, J. *et al.* Genome-wide analysis in UK Biobank identifies four loci associated with mood instability and genetic correlation with major depressive disorder, anxiety disorder and schizophrenia. *Transl. Psychiatry* **7,** 1264 (2017).
83. Reus, L. M. *et al.* Association of polygenic risk for major psychiatric illness with subcortical volumes and white matter integrity in UK Biobank. *Sci. Rep.* **7,** 42140 (2017).
84. Howard, D. M. *et al.* Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank. *Transl. Psychiatry* **7,** 1263 (2017).
85. Howard, D. M. *et al.* Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat. Commun.* **9,** 1470 (2018).
86. Hall, L. S. *et al.* Genome-wide meta-analyses of stratified depression in Generation Scotland and UK Biobank. *Transl. Psychiatry* **8,** 9 (2018).
87. Deary, V. *et al.* Genetic contributions to self-reported tiredness. *Mol. Psychiatry* **23,** 609–620 (2018).
88. Hagenaars, S. P. *et al.* Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Mol. Psychiatry* **21,** 1624–1632 (2016).
89. Rutten-Jacobs, L. C. A. *et al.* Genetic Study of White Matter Integrity in UK Biobank (N=8448) and the Overlap With Stroke, Depression, and Dementia. *Stroke* STROKEAHA.118.020811 (2018). doi:10.1161/STROKEAHA.118.020811
90. Smeland, O. B. *et al.* Identification of Genetic Loci Jointly Influencing Schizophrenia Risk and the Cognitive Traits of Verbal-Numerical Reasoning, Reaction Time, and General Cognitive Function. *JAMA Psychiatry* **74,** 1065 (2017).
91. Croarkin, P. E. *et al.* Genetic Risk Score Analysis in Early-Onset Bipolar Disorder. *J. Clin. Psychiatry* **78,** 1337–1343 (2017).
92. McElroy, S. L. *et al.* Bipolar disorder with binge eating behavior: a genome-wide association study implicates PRR5-ARHGAP8. *Transl. Psychiatry* **8,** 40 (2018).
93. Clarke, T.-K. *et al.* Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). *Mol. Psychiatry* **22,** 1376–1384 (2017).
94. Warren, H. R. *et al.* Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.* **49,** 403–415 (2017).
95. Klarin, D. *et al.* Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat. Genet.* **49,** 1392–1397 (2017).

96. Chan, K. *et al.* Genetic Variation at the ADAMTS7 Locus is Associated With Reduced Severity of Coronary Artery Disease. *J. Am. Heart Assoc.* **6,** e006928 (2017).

97. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49,** 1385–1391 (2017).

98. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery DiseaseNovelty and Significance. *Circ. Res.* **122,** 433–443 (2018).

99. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case–control association mapping by proxy using family history of disease. *Nat. Genet.* **49,** 325–331 (2017).

100. Lyall, D. M. *et al.* Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank. *JAMA Cardiol.* **2,** 882 (2017).

101. Li, X. *et al.* MR-PheWAS: exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK Biobank. *Ann. Rheum. Dis.* annrheumdis-2017-212534 (2018). doi:10.1136/annrheumdis-2017-212534

102. Tikkanen, E. *et al.* Biological Insights Into Muscular Strength: Genetic Findings in the UK Biobank. *Sci. Rep.* **8,** 6451 (2018).

103. Ritchie, M. D. *et al.* Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *Am. J. Hum. Genet.* **86,** 560–572 (2010).

104. Thériault, S. *et al.* A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. *Nat. Commun.* **9,** 988 (2018).

105. Emdin, C. A. *et al.* Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. *JAMA* **317,** 626 (2017).

106. van Zon, S. K. R. *et al.* The interaction of genetic predisposition and socioeconomic position with type 2 diabetes mellitus. *Psychosom. Med.* 1 (2018). doi:10.1097/PSY.0000000000000562

107. Zhao, W. *et al.* Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* **49,** 1450–1457 (2017).

108. Tachmazidou, I. *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. *Am. J. Hum. Genet.* **100,** 865–884 (2017).

109. Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLOS Genet.* **13,** e1006977 (2017).

110. Gill, D. *et al.* Age at menarche and adult body mass index: a Mendelian randomization study. *Int. J. Obes.* (2018). doi:10.1038/s41366-018-0048-7

111. Turcot, V. *et al.* Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat. Genet.* **50,** 26–41 (2018).

112. Astley, C. M. *et al.* Genetic Evidence That Carbohydrate-Stimulated Insulin Secretion Leads to Obesity. *Clin. Chem.* **64,** 192–200 (2018).

113. Zengini, E. *et al.* Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nat. Genet.* **50,** 549–558 (2018).

114. Cronin, R. M. *et al.* Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Front. Genet.* **5,** 1–14 (2014).

115. Paré, G., Mao, S. & Deng, W. Q. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci. Rep.* **7,** 12665 (2017).

116. Márquez-Luna, C., Loh, P.-R. & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41,** 811–823 (2017).

117. Beaumont, R. N. *et al.* Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics. *Hum. Mol. Genet.* **27,** 742–756 (2018).

118. Tyrrell, J. *et al.* Gene–obesogenic environment interactions in the UK Biobank study. *Int. J. Epidemiol.* 559–575 (2017). doi:10.1093/ije/dyw337

119. Usher-Smith, J. A. *et al.* External validation of risk prediction models for incident colorectal cancer using UK Biobank. *Br. J. Cancer* **118,** 750–759 (2018).
120. Abana, C. O. *et al.* IL-6 variant is associated with metastasis in breast cancer patients. *PLoS One* **12,** e0181725 (2017).
121. Hoffman, J. D. *et al.* Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. *PLOS Genet.* **13,** e1006690 (2017).
122. van Duijnhoven, F. J. B. *et al.* Circulating concentrations of vitamin D in relation to pancreatic cancer risk in European populations. *Int. J. Cancer* **142,** 1189–1201 (2018).
123. Taylor, M. *et al.* Is smoking heaviness causally associated with alcohol use? A Mendelian randomization study in four European cohorts. *Int. J. Epidemiol.* 1–8 (2018). doi:10.1093/ije/dyy027
124. Wain, L. V *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat. Genet.* **49,** 416–425 (2017).
125. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26,** 1205–1210 (2010).
126. Liao, K. P. *et al.* Phenome-Wide Association Study of Autoantibodies to Citrullinated and Noncitrullinated Epitopes in Rheumatoid Arthritis. *Arthritis Rheumatol.* (2017). doi:10.1002/art.39974
127. Neuraz, A. *et al.* Phenome-Wide Association Studies on a Quantitative Trait: Application to TPMT Enzyme Activity and Thiopurine Therapy in Pharmacogenomics. *PLoS Comput. Biol.* **9,** (2013).
128. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. Advantages and pitfalls in the application of mized model association methods. *Nat. Genet.* **46,** 100–106 (2014).
129. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am. J. Hum. Genet.* **101,** 37–49 (2017).
130. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* (2018).
131. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).
132. Lee, J. J. & Chow, C. C. LD score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *bioRXiv* 1–24 (2017).
133. Meng, W. *et al.* A Genome-Wide Association Study Finds Genetic Associations with Broadly-Defined Headache in UK Biobank (N = 223,773). *EBioMedicine* **28,** 180–186 (2018).
134. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
135. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLOS Genet.* **13,** e1006711 (2017).
136. Yang, J. *et al.* A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res. Hum. Genet.* **13,** 517–524 (2010).
137. Bastarache, L. *et al.* Phenotype risk scores identify pations with unrecognized Mendelian disease patterns. *Science (80-.).* **359,** 1233–1239 (2018).
138. Burgess, S., Timpson, N. J., Ebrahim, S. & Smith, G. D. Mendelian randomization: Where are we now and where are we going? *Int. J. Epidemiol.* **44,** 379–388 (2015).
139. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30,** 2375–2376 (2014).
140. Dobrzykowski, D. D. Examining heterogeneous patterns of electronic health records use: a contingency perspective and assessment. *Int. J. Healthc. Inf. Syst. Informatics* **7,** 1–14 (2012).
141. Au Yeung, S. L. *et al.* Aldehyde dehydrogenase 2—a potential genetic risk factor for lung function among southern Chinese: evidence from the Guangzhou Biobank Cohort Study. *Ann. Epidemiol.* **24,** 606–611 (2014).

142. Kuhnert, R. *et al.* A modified self-controlled case series method to examine association between multidose vaccinations and death. *Stat. Med.* **30,** 666–677 (2011).
143. Zhou, X., Douglas, I. J., Shen, R. & Bate, A. Signal Detection for Recently Approved Products: Adapting and Evaluating Self-Controlled Case Series Method Using a US Claims and UK Electronic Medical Records Database. *Drug Saf.* **41,** 523–536 (2018).
144. Schumie, M. J., Trifiro, G., Coloma, P. M., Ryan, P. B. & Madigan, D. Detecting Adverse drug reactions following long-term exposure in longitudinal observational data: the exposure-adjusted self-controlled case series. *Stat. Methods Med. Res.* **25,** 2577–2592 (2016).
145. Maclure, M. *et al.* When should case-only designs be used for safety monitoring of medical products? *Pharmacoepidemiol. Drug Saf.* (2012). doi:10.1002/pds.2330
146. Simpson, S. E. *et al.* Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics* (2013). doi:10.1111/biom.12078
147. Petersen, I., Douglas, I. & Whitaker, H. Self controlled case series methods: an alternative to standard epidemiological study designs. *BMJ* **354,** i4515 (2016).
148. Madigan, D., Ryan, P. B. & Schuemie, M. J. Does design matter? Systematic evaluation of the impact of analytical choices on effect esitmates in observational studies. *Ther. Adv. Drug Saf.* **4,** 53–62 (2013).
149. Haneuse, S., Chan, H. T. H. & Daniels, M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why? doi:10.13063/2327-9214.1203
150. Haneuse, S. *et al.* Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology* **32,** 229–239 (2009).
151. Zheng, K., Gao, J., Ngiam, K. Y., Ooi, B. C. & Yip, W. L. J. Resolving the Bias in Electronic Medical Records. *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '17* 2171–2180 (2017). doi:10.1145/3097983.3098149
152. Phelan, M., Bhavsar, N. & Goldstein, B. A. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. *eGEMs (Generating Evid. Methods to Improv. patient outcomes)* **5,** 22 (2017).
153. Goldstein, B. A., Bhavsar, N. A., Phelan, M. & Pencina, M. J. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am. J. Epidemiol.* **184,** 847–855 (2016).
154. Rusanov, A., Weiskopf, N. G., Wang, S. & Weng, C. Hidden in plain sight: Bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med. Inform. Decis. Mak.* **14,** 1–9 (2014).
155. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32,** 1–22 (2003).
156. Avery, C. L., Monda, K. L. & North, K. E. Genetic association studies and the effect of misclassification and selection bias in putative confounders. *BMC Proc.* **3,** S48 (2009).
157. Haneuse, S. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Med. Care* **54,** 1–16 (2017).
158. Carnegie, N. B., Harada, M. & Hill, J. L. Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder. *J. Res. Educ. Eff.* **9,** 395–420 (2016).
159. Uddin, M. J. *et al.* Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int. J. Clin. Pharm.* **38,** 714–723 (2016).
160. Zhang, X., Faries, D. E., Li, H., Stamey, J. D. & Imbens, G. W. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiol. Drug Saf.* **27,** 373–382 (2018).
161. Norén, G. N. *et al.* Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: Lessons for developing a risk identification and analysis system. *Drug Saf.* (2013). doi:10.1007/s40264-013-0095-x
162. Schuemie, M. J., Ryan, P. B., Dumouchel, W., Suchard, M. A. & Madigan, D. Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Stat. Med.* **33,** 209–

218 (2014).

163. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D. & Suchard, M. A. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc. Natl. Acad. Sci.* **115,** 2571–2577 (2018).

164. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

165. Sun, Z., Mukherjee, B., Estes, J. P., Vokonas, P. S. & Park, S. K. Exposure enriched outcome dependent designs for longitudinal studies of gene–environment interaction. *Stat. Med.* **36,** 2947–2960 (2017).

166. Schildcrout, J. S., Rathouz, P. J., Zelnick, L. R., Garbett, S. P. & Heagerty, P. J. Biased sampling designs to improve research efficiency: Factors influencing pulmonary function over time in children with asthma. *Ann. Appl. Stat.* **9,** 731–753 (2015).

167. Schildcrout, J. S., Schisterman, E. F., Mercaldo, N. D., Rathouz, P. J. & Heagerty, P. J. Extending the case-control design to longitudinal data: stratified sampling based on repeated binary outcomes. *Epidemiology* **29,** 67–75 (2018).

168. ICD Code Informational Website. Available at: https://www.cdc.gov/nchs/icd/index.htm.

169. Pendergrass, S. A. & Ritchie, M. D. Phenome-wide association studies: leveraging comprehensive phenotypic and genotypic data for discovery. *Curr. Genet. Med. Rep.* **42,** 407–420 (2016).

170. eMERGE PheKB Website. Available at: https://phekb.org.

171. Liao, K. P. *et al.* Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One* (2015). doi:10.1371/journal.pone.0136651

172. Liao, K. P. *et al.* Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* (2015). doi:10.1136/bmj.h1885

173. Ananthakrishnan, A. N. *et al.* Identification of Nonresponse to Treatment Using Narrative Data in an Electronic Health Record Inflammatory Bowel Disease Cohort. *Inflamm. Bowel Dis.* (2016). doi:10.1097/MIB.0000000000000580

174. Castro, V. *et al.* Identification of subjects with polycystic ovary syndrome using electronic health records. *Reprod. Biol. Endocrinol.* (2015). doi:10.1186/s12958-015-0115-z

175. McCoy, T. H. *et al.* Genome-wide Association Study of Dimensional Psychopathology Using Electronic Health Records. *Biol. Psychiatry* (2018). doi:10.1016/j.biopsych.2017.12.004

176. Sinnott, J. A. *et al.* Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Hum. Genet.* **133,** 1369–1382 (2014).

177. Yu, S. *et al.* Surrogate-assisted feature extraction for high-throughput phenotyping. *J. Am. Med. Inform. Assoc.* **24,** e143–e149 (2017).

178. Yu, S. *et al.* Enabling phenotypic big data with PheNorm. *J. Am. Med. Informatics Assoc.* **25,** 54–60 (2018).

179. Castro, V. M. *et al.* Large-scale identification of patients with cerebral aneurysms using natural language processing. *Neurology* **88,** 164–168 (2017).

180. Kermany, D. S. *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172,** 1122–1131.e9 (2018).

181. Teixeira, P. L. *et al.* Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J. Am. Med. Informatics Assoc.* **24,** 162–171 (2017).

182. Gan, M., Li, W., Zeng, W., Wang, X. & Jiang, R. Mimvec: a deep learning approach for analyzing the human phenome. *BMC Syst. Biol.* **11,** 76 (2017).

183. Lange, J. M., Hubbard, R. A., Inoue, L. Y. T. & Minin, V. N. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics* (2015). doi:10.1111/biom.12252

184. Bergeron, P. J., Asgharian, M. & Wolfson, D. B. Covariate bias induced by length-biased sampling of failure times. *J. Am. Stat. Assoc.* **103,** 737–742 (2008).

185. Castro, V. M. *et al.* Validation of electronic health record phenotyping of bipolar disorder cases

and controls. *Am. J. Psychiatry* **172,** 363–372 (2015).
186. Baiardini, I., Braido, F., Bonini, M., Compalati, E. & Canonica, G. W. Why do doctors and patients not follow guidelines? *Curr. Opin. Allergy Clin. Immunol.* **9,** 228–233 (2009).
187. Hubbard, R. A. *et al.* Classification accuracy of claims-based methods for identifying providers failing to meet performance targets. *Stat. Med.* (2015). doi:10.1002/sim.6318
188. Huang, J. *et al.* PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *J. Am. Med. Informatics Assoc.* (2018). doi:10.1093/jamia/ocx137
189. Duffy, S. W. *et al.* A simple model for potential use with a misclassified binary outcome in epidemiology. *J. Epidemiol. Community Health* **58,** 712–717 (2004).
190. Tsoi, L. C. *et al.* Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nat. Commun.* **8,** 1–8 (2017).
191. Mersha, T. B. & Abebe, T. Self-reported race/ethnicity in the age of genomic research: Its potential impact on understanding health disparities. *Hum. Genomics* **9,** 1–15 (2015).
192. Banda, Y. *et al.* Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. *Genetics* **200,** 1285–1295 (2015).
193. Spangler, L. *et al.* A comparison of self-reported oral contraceptive use and automated pharmacy data in perimenopausal and early postmenopausal women. *Ann. Epidemiol.* (2015). doi:10.1016/j.annepidem.2014.10.009
194. Wells, B. J., Chagin, K. M., Nowacki, A. S. & Kattan, M. W. Strategies for handling missing data in electronic health record derived data. *EGEMS (Washington, DC)* **1,** 1035 (2013).
195. Hormozdiari, F. *et al.* Imputing Phenotypes for Genome-wide Association Studies. *Am. J. Hum. Genet.* **99,** 89–103 (2016).
196. Beaulieu-Jones, B. K. & Moore, J. H. Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders. *Biocomput. 2017* 207–218 (2017). doi:10.1142/9789813207813_0021
197. Beaulieu-Jones, B. K. *et al.* Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Med. Informatics* **6,** e11 (2018).
198. Good, P. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. (Springer, 2005).
199. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32,** 361–369 (2008).
200. Han, B., Kang, H. M. & Eskin, E. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.* **5,** 1–13 (2009).
201. Lin, D. Y. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21,** 781–787 (2005).
202. Seaman, S. R. & Müller-Myhsok, B. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.* **76,** 399–408 (2005).
203. Duggal, P., Gillanders, E. M., Holmes, T. N. & Bailey-Wilson, J. E. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* **9,** 1–8 (2008).
204. Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A. & Winkler, C. A. Accounting for multiple comparisons in a genome- wide association study (GWAS). 0–20 (2010). doi:10.1186/1471-2164-11-724
205. Zhang, X., Huang, S., Sun, W. & Wang, W. Rapid and robust resampling-based multiple-testing correction with application in a genome-wide expression quantitative trait loci study. *Genetics* **190,** 1511–1520 (2012).
206. Li, M. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131,** 747–756 (2012).
207. Thompson, S. G. Systematic Review: Why sources of heterogeneity in meta-analysis should be

investigated. *Bmj* **309,** 1351–1355 (1994).
208. Fletcher, J. What is heterogeneity and is it important? *British Medical Journal* (2007). doi:10.1136/bmj.39057.406644.68
209. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses Need for consistency. *BMJ* **327,** 557–560 (2003).
210. Kriston, L. Dealing with clinical heterogeneity in meta-analysis. Assumptions, methods, interpretation. *Int. J. Methods Psychiatr. Res.* **22,** 1–15 (2013).
211. Li, Y. & Ghosh, D. Assumption weighting for incorporating heterogeneity into meta-analysis of genomic data. *Bioinformatics* **28,** 807–814 (2012).
212. Grimmer, J., Messing, S. & Westwood, S. J. Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods. *Polit. Anal.* **25,** 413–434 (2017).
213. Gagnier, J. J., Moher, D., Boon, H., Bombardier, C. & Beyene, J. An empirical study using permutation-based resampling in meta-regression. *Syst. Rev.* **1,** 1–9 (2012).
214. Chubak, J., Onega, T., Zhu, W., Buist, D. S. M. & Hubbard, R. A. An Electronic Health Record-based Algorithm to Ascertain the Date of Second Breast Cancer Events. *Med. Care* (2017). doi:10.1097/MLR.0000000000000352
215. Manrai, A. K. *et al.* Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health. *Annu. Rev. Public Heal.* **38,** 279–94 (2017).
216. Fan, J. W., Li, J. & Lussier, Y. A. Semantic Modeling for Exposomics with Exploratory Evaluation in Clinical Context. *J. Healthc. Eng.* (2017). doi:10.1155/2017/3818302
217. Baek, J. *et al.* Methods to study variation in associations between food store availability and body mass in the multi-ethnic study of atherosclerosis. *Epidemiology* **28,** 403–411 (2017).
218. Bazemore, A. W. *et al.* 'Community vital signs': Incorporating geocoded social determinants into electronic records to promote patient and population health. *J. Am. Med. Informatics Assoc.* **23,** 407–412 (2016).
219. Christine, P. J. *et al.* Exposure to Neighborhood Foreclosures and Changes in Cardiometabolic Health: Results from MESA. *Am. J. Epidemiol.* **185,** 106–114 (2017).
220. Frederickson Comer, K., Grannis, S., Dixon, B. E., Bodenhamer, D. J. & Wiehe, S. E. Incorporating Geospatial Capacity within Clinical Data Systems to Address Social Determinants of Health. *Public Health Rep.* **3,** 54–61 (2011).
221. Sánchez, B. N., Sanchez-Vaznaugh, E. V., Uscilka, A., Baek, J. & Zhang, L. Differential associations between the food environment near schools and childhood overweight across race/ethnicity, gender, and grade. *Am. J. Epidemiol.* **175,** 1284–1293 (2012).
222. Xie, S., Greenblatt, R., Levy, M. Z. & Himes, B. E. Enhancing Electronic Health Record Data with Geospatial Information. in *AMIA Jt Summits Transl Sci Proc* 123–132 (2017).
223. Al-Azwani, I. K. & Aziz, H. A. Integration of Wearable Technologies into Patients' Electronic Medical Records. *Qual. Prim. Care* **24,** 151–155 (2016).
224. Polzer, N. & Gewald, H. A Structured Analysis of Smartphone Applications to Early Diagnose Alzheimer´s Disease or Dementia. *Procedia Comput. Sci.* **113,** 448–453 (2017).
225. Norén, G. N., Hopstadius, J., Bate, A., Star, K. & Edwards, I. R. Temporal pattern discovery in longitudinal electronic patient records. *Data Min. Knowl. Discov.* (2010). doi:10.1007/s10618-009-0152-3
226. Boland, M. R., Shahn, Z., Madigan, D., Hripcsak, G. & Tatonetti, N. P. Birth month affects lifetime disease risk: A phenome-wide method. *J. Am. Med. Informatics Assoc.* (2015). doi:10.1093/jamia/ocv046
227. Liu, M. *et al.* Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *J Am Med Inf. Assoc* **20,** 420–426 (2013).
228. Peterson, J. F. *et al.* Electronic Health Record Design and Implementation for Pharmacogenomics: a Local Perspective HHS Public Access. *Genet Med* **15109,** 833–841 (2013).
229. Madigan, D. & Shin, J. Drospirenone-containing oral contraceptives and venous thromboembolism: an analysis of the FAERS database. *Open Access J. Contracept.* **9,** 29–32

(2018).
230. Shuldiner, A. R. *et al.* The pharmacogenomics research network translational pharmacogenetics program: Overcoming challenges of real-world implementation. *Clin. Pharmacol. Ther.* **94,** 207–210 (2013).
231. Kuang, Z. *et al.* Computational Drug Repositioning Using Continuous Self- Controlled Case Series. 491–500 (2017). doi:10.1145/2939672.2939715
232. Paige, E. *et al.* Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *Am. J. Epidemiol.* **187,** 1530–1538 (2018).
233. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. P. A. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Informatics Assoc.* **24,** 198–208 (2017).
234. Caballero, K. & Akella, R. Dynamic Estimation of the Probability of Patient Readmission to the ICU using Electronic Medical Records. *AMIA Annu. Symp. Proc.* **2015,** 1831–40 (2015).
235. Aczon, M. *et al.* Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks. 1–18 (2017).
236. Rose, S. Machine Learning for Prediction in Electronic Health Data. *JAMA Netw. Open* **1,** 1–3 (2018).
237. Steorts, R. C., Hall, R. & Fienberg, S. E. A Bayesian Approach to Graphical Record Linkage and Deduplication. *J. Am. Stat. Assoc.* **111,** 1660–1672 (2016).
238. Sayers, A., Ben-Shlomo, Y., Blom, A. W. & Steele, F. Probabilistic record linkage. *Int. J. Epidemiol.* **45,** 954–964 (2016).
239. Vatsalan, D., Christen, P. & Verykios, V. S. A taxonomy of privacy-preserving record linkage techniques. *Inf. Syst.* **38,** 946–969 (2013).
240. Mamun, A. Al, Aseltine, R. & Rajasekaran, S. Efficient record linkage algorithms using complete linkage clustering. *PLoS One* **11,** 1–21 (2016).
241. Schmidlin, K., Clough-Gorr, K. M. & Spoerri, A. Privacy Preserving Probabilistic Record Linkage (P3RL): A novel method for linking existing health-related data and maintaining participant confidentiality. *BMC Med. Res. Methodol.* **15,** 1–10 (2015).
242. Long, Q. Statistical Methods for Handling Missing Data in Distributed Health Data Networks. in *Joint Statistical Meetings* (2018).
243. Tang, L., Zhou, L. & Song, P. X.-K. *Method of Divide-and-Combine in Regularised Generalised Linear Models for Big Data*. (2016).
244. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44,** 1–22 (2013).
245. Hripcsak, G. *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc. Natl. Acad. Sci.* (2016). doi:10.1073/pnas.1510502113
246. Simon, G. E., Peterson, D. & Hubbard, R. Is treatment adherence consistent across time, across different treatments and across diagnoses? *Gen. Hosp. Psychiatry* (2013). doi:10.1016/j.genhosppsych.2012.10.001
247. Delaney, J. T. *et al.* Predicting Clopidogrel Response Using DNA Samples Linked to an Electronic Health Record. *Clin. Pharmacol. Ther.* **91,** 257–263 (2012).
248. Hernán, M. A. *et al.* Observationals studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19,** 766–779 (2008).
249. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am. J. Epidemiol.* **183,** 758–764 (2016).
250. Danaei, G., García Rodríguez, L. A., Cantero, O. F., Logan, R. W. & Hernán, M. A. Electronic medical records can be used to emulate target trials of sustained treatment strategies. *J. Clin. Epidemiol.* **96,** 12–22 (2018).
251. Bolland, M. J., Grey, A., Gamble, G. D., Reid, I. R. & Coleman, W. B. Concordance of results

from randomized and observational analyses within the same study: A re-analysis of the women's health initiative limited-access dataset. *PLoS One* **10,** (2015).

252. Franklin, J. M. & Schneeweiss, S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? *Clin. Pharmacol. Ther.* **102,** 924–933 (2017).

doi:10.20944/preprints201809.0388.v1

Figures

Figure 1: Overall Distribution of Selected Biobank-Based Publications by Year and Type

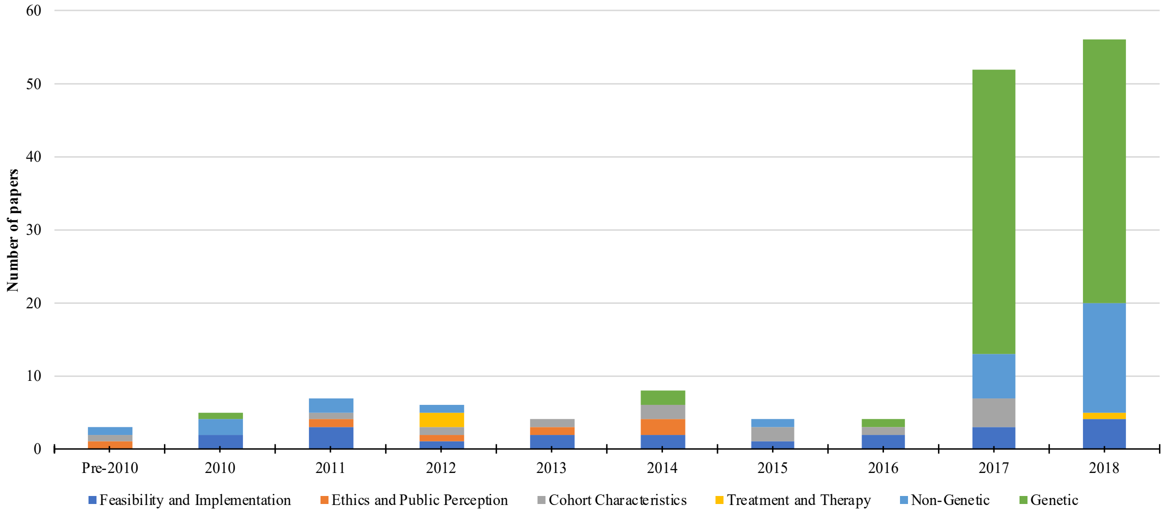


Figure 2: Boxplots of Ratio of PheWAS Code Prevalence in MGI vs. UK Biobank Across Phenome

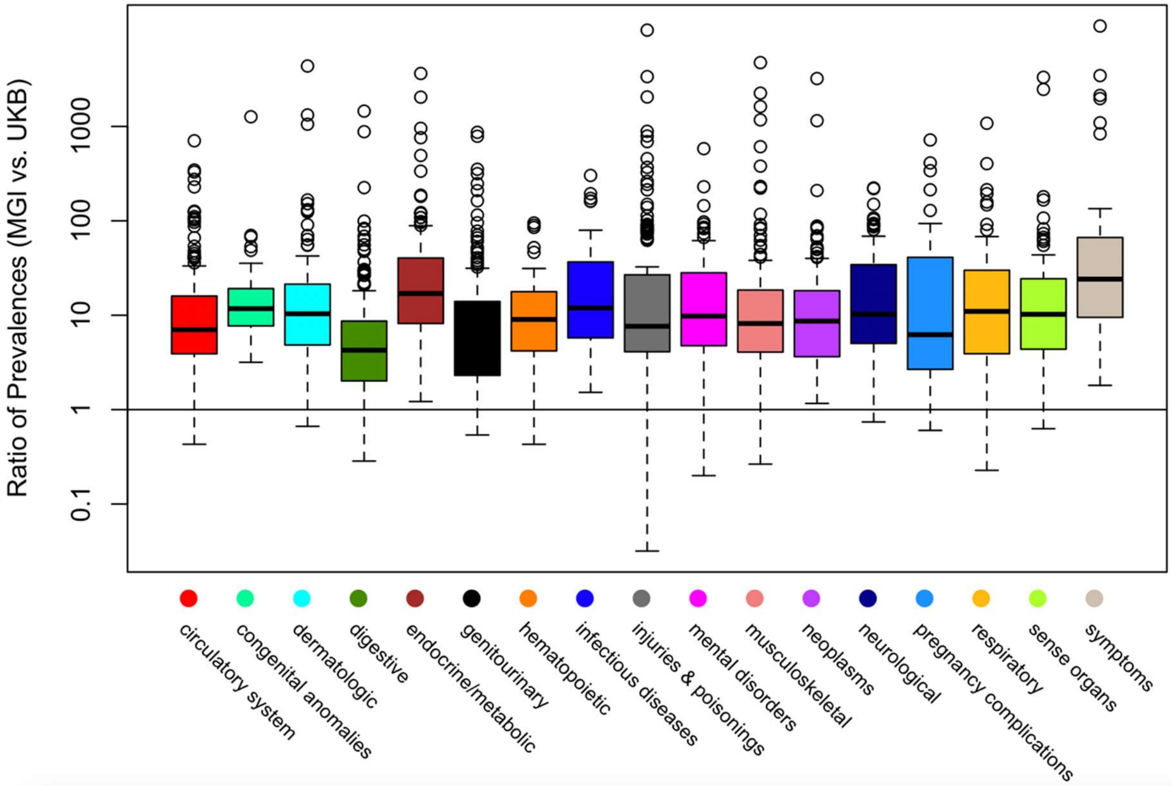


Figure 3: Potential Data Sources for Generating the Phenome

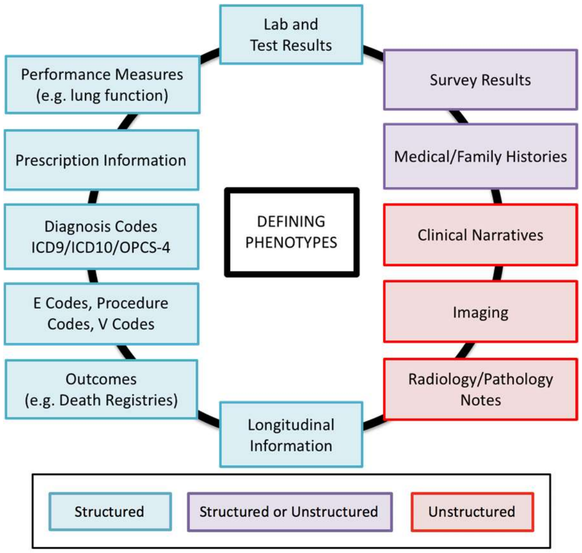

Lab and Test Results
Survey Results
Medical/Family Histories
Clinical Narratives
Imaging
Radiology/Pathology Notes
Longitudinal Information
Outcomes (e.g. Death Registries)
E Codes, Procedure Codes, V Codes
Diagnosis Codes ICD9/ICD10/OPCS-4
Prescription Information
Performance Measures (e.g. lung function)
DEFINING PHENOTYPES
Structured
Structured or Unstructured
Unstructured

Figure 4: Relationship between True and Observed Phenome

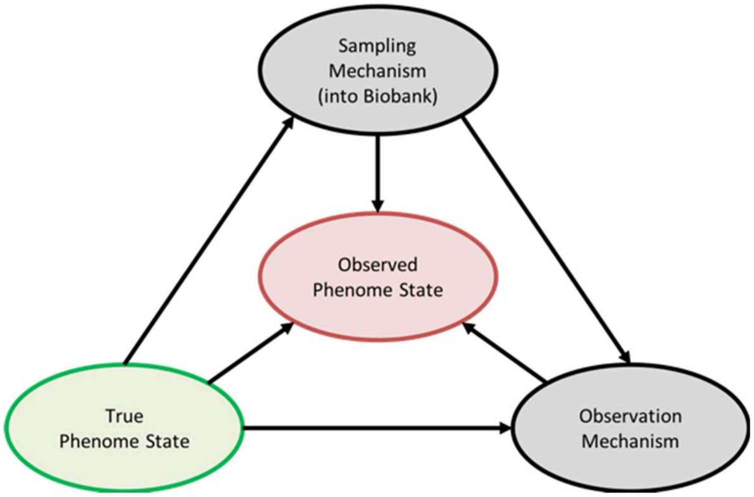
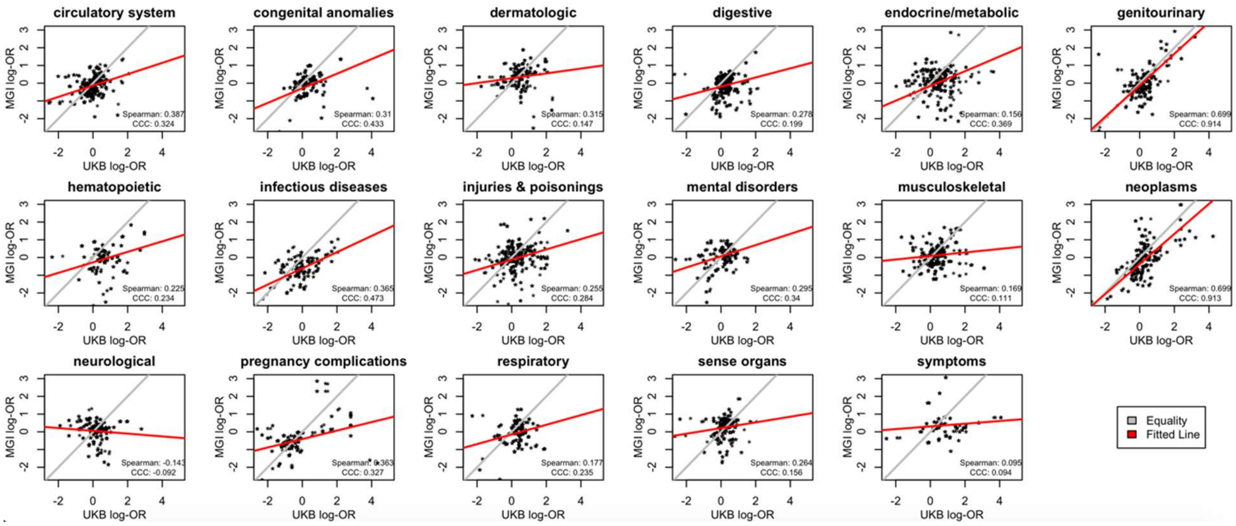
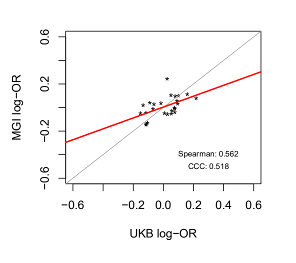
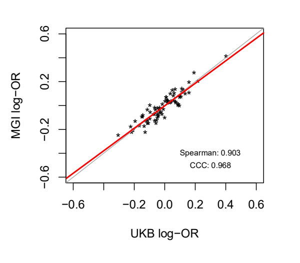
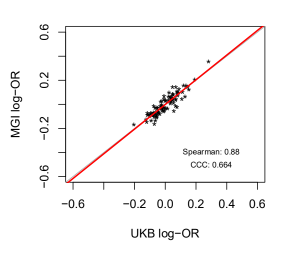
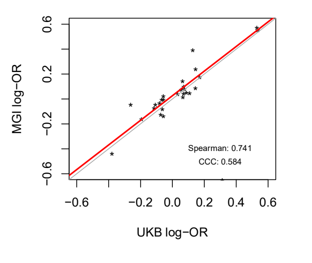


Figure 5: Log-Odds Ratios of having Breast Cancer Diagnosis by Other Phenotype Diagnoses*



* Each point represents a phenotype in MGI and UK Biobank in a particular disease category (say respiratory) and the corresponding cross-classified log-odds ratios capturing the association between breast cancer diagnosis and diagnosis of the other phenotype in MGI and UK Biobank. 2,025 subjects had observed breast cancer in MGI and 12,680 subjects had breast cancer in UK Biobank. The two lines correspond to equality of the estimates and a fitted line to the points. "Spearman" indicates the Spearman correlation and "CCC" indicates Lin's concordance correlation coefficient, which is a measure of agreement (with 1 being perfect agreement).

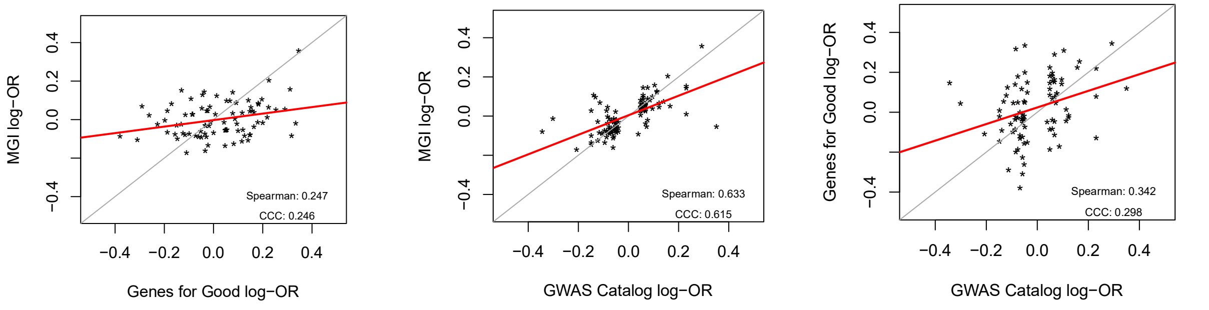
Figure 6: A Comparison of GWAS Results in MGI and UK Biobank (UKB) for Selected Cancer Phenotypes*

Colorectal Cancer	Prostate Cancer	Breast Cancer	Melanoma
			

* Each point represents a SNP identified as being related to the corresponding phenotype in the NHGRI-EBI GWAS catalog. The point location corresponds to the log-odds ratio association between the SNP and the phenotype of interest in MGI and UK Biobank. The two lines correspond to equality of the estimates and a fitted line to the points (excluding any outlying points with absolute log-OR greater than 0.6). "Spearman" indicates the Spearman correlation and "CCC" indicates Lin's concordance correlation coefficient, which is a measure of agreement (with 1 being perfect agreement).

Figure 7: A Comparison of Breast Cancer GWAS Results in MGI with GFG*

MGI and GFG	MGI and GWAS Catalog	GFG and GWAS Catalog



Potential Explanation for Differences:
There were 2,025 female breast cancers in MGI out of 16,297 women (12.4%). However, there were only 115 female breast cancers in GFG out of 10,802 women (1.1%). This is explained by the different age distributions, since GFG subjects were younger on average, with a mean age of 36.9 years in GFG and 54.2 years in MGI. Therefore, many GFG subjects are not in the age window of susceptibility for breast cancer, which is a disease more common after 50. Differences in the log-OR estimates can also be partly explained by differences in sample sizes, leading to GFG estimates that are much more variable than estimates in MGI or those reported in the GWAS Catalog. Additionally, differences in log-OR estimates may result from different phenotype definitions.

* Each point represents a SNP identified as being related to the corresponding phenotype in the NHGRI-EBI GWAS catalog and the corresponding estimated log-OR SNP-phenotype associations in MGI, GFG, or reported in the GWAS catalog. The two lines correspond to equality of the estimates and a fitted line to the points (excluding any outlying points with absolute log-OR greater than 0.6). "Spearman" indicates the Spearman correlation and "CCC" indicates Lin's concordance correlation coefficient, which is a measure of agreement (with 1 being perfect agreement).

Tables

Table 1: Description of Selected Major Biobanks

Biobank	Start year	Location	Age	Size	Type*	Institution	Access	Connected to EHR	Linked with prescriptions?	Linked to death registry?	Biospecimen Collected	Survey	Website
All of Us	2018	USA	18+	1 million (goal)	Health system	National Institutes of Health	Not yet available	Yes	Yes**	-	Blood, saliva, urine	Yes	https://www.joinallofus.org/en
BioBank Japan	2003	Japan	-	200,000+	Health system	Ministry of Education, Culture, Sports, Science and Technology	Inquire with biobank	Yes	-	Yes	Blood (buccal swabs or nail/hair trimmings)	Yes	http://www.pgrn.org/biobank-japan.html
BioME	-	Mount Sinai Health System	-	42,000+	Health system	Mount Sinai Health System	Inquire with biobank	Yes	-	-	Blood	Yes	https://icahn.mssm.edu/research/ipm/programs/biome-biobank
BioVU	2007	Tennessee	18+	250,000+	Health system	Vanderbilt University	Inquire with biobank	Yes	No	No	Blood	No	https://victr.vanderbilt.edu/pub/biovu/?sid=194
China Kadoorie Biobank	2004	China	30-79	510,000+	Population	University of Oxford + Chinese Academy of Medical Sciences	Application for researchers	Yes	-	Yes	Blood	Yes	http://www.ckbiobank.org/site/
deCODE Genetics	1996	Iceland	-	~500,000	Commercial	deCODE (Amgen)	Inquire with biobank	Yes	-	-	-	-	https://www.decode.com/
DiscovEHR	2014	Geisinger Health System; Regeneron Genetics Center	18+	50000	Health system	Regeneron Genetics Center + Geisinger Health System	Inquire with biobank	Yes	No	No	Blood	No	http://www.discovehrshare.com/
eMERGE Network	2007	NHGRI	All	126,000+	Network of biobanks	National Human Genome Research Institute	Application for researchers	Yes	No	No	Genetic results obtained from external sources	No	https://emerge.mc.vanderbilt.edu/
Generation Scotland	2006	Scotland	18-65	30,000+	Population	University of Edinburgh	Application for researchers	Yes	Yes	Yes	Blood, urine (saliva for some subjects)	Yes	https://www.ed.ac.uk/generation-scotland
Guangzhou Biobank Cohort Study	2003	Guangzhou	50+	~30,000	Population	Universities of Birmingham and Hong Kong + The Guangzhou Occupational Diseases Prevention and Treatment Center	Inquire with biobank	Yes	No	Yes	Blood	Yes	https://www.birmingham.ac.uk/research/activity/mds/projects/HaPS/PHEB/Guangzhou/index.aspx
HUNT - Nord-Trøndelag Health Study	2002	Nord-Trøndelag County, Norway	20+	125,000	Population	Norwegian University of Science and Technology	Application for researchers	Yes	Yes	Yes	Blood (urine for some subjects)	Yes	https://www.ntnu.edu/hunt/hunt-biobank
Kaiser Permanente Research Bank	2008	Kaiser Permanente	18+	308,425	Health system	Kaiser Permanente	Application for researchers	Yes	Yes	-	Blood, saliva	Yes	https://researchbank.kaiserpermanente.org/
Michigan Genomics Initiative	2012	Michigan	18+	60,000+	Health system	University of Michigan	Inquire with biobank	Yes	No	Yes**	Blood	Yes*	https://www.michigangenomics.org
Million Veterans Program	2011	USA	-	600,000+	Health system	US Dept. of Veterans Affairs	Inquire with biobank	Yes	-	-	Blood	Yes	https://www.research.va.gov/mvp/
MyCode Community Health Initiaitve (Geisinger)	2007	Geisinger Health System	7+	190,000+	Health System	Geisinger Health	Inquire with biobank	Yes	No	No	Blood or saliva	No	https://www.geisinger.org/mycode#cgg
Partners HealthCare Biobank	2010	Brigham and Women's Hospital; Massachusetts General	18+	80,000+	Health System	Partners Healthcare	Inquire with biobank	Yes	No	No	Blood	Yes	https://biobank.partners.org/
UK Biobank	2006	United Kingdom	40-69	500,000	Population	UK Biobank charity	Application for researchers	Yes	-	-	Blood, urine, saliva	Yes	http://www.ukbiobank.ac.uk/about-biobank-uk/
CARTaGENE	2009	Quebec	40-69	43,000	Population	CHU Sainte-Justine Research Center	Application for researchers	No	No	Yes	Blood, urine	Yes	https://www.cartagene.qc.ca/en/about
Genes for Good	2015	USA	18+	77,700+	Self-initiated	University of Michigan	Inquire with biobank	No	No	No	Saliva	Yes	https://genesforgood.sph.umich.edu
Trans-Omics for Precision Medicine (TopMed)	2014	USA (various sites)	-	~145,000	Consortium of studies	University of Washington	NIH Database of Genotypes and Phenotypes (dbGaP)	No	No	No	Genetic results obtained from external sources	No	https://www.nhlbiwgs.org/
Lifelines	2006	Northern Netherlands	All	167000+	Population	Lifelines Biobank	Application for researchers	-	No	No	Blood, urine	Yes	https://www.lifelines.nl/researcher

- indicates information is unknown; * we chose categories we thought best fit each biobank; ** indicates we found a source saying the resource is being developed or will be available in the future

Note: The information in this table is ascertained to the best of our knowledge. Where it indicates 'yes', this means we were able to find a source that indicates this is a feature of the biobank. Where it indicates 'no', this means that it was either absent or there was sufficient reason to believe the resource is unavailable at the biobank. It is best to contact the biobank to confirm the availability of resources that are unknown or indicated as not available.

Table 2: Prevalences of Selected Conditions in the Michigan Genomics Initiative, UK Biobank, and Genes for Good along with Estimates from their Respective National Populations∇

	MGI **(Academic Medical Center)**	**GFG** **(Subject-Initiated)**	**US**	**UKB** **(Population-Based)**	**UK**
	N = 30,702	N = 15,156		N = 408,961	
Psychiatric/Neurologic					
Depression	21.7 (6,651)	70.2 (10,642)	16.9**	2.9 (11,918)	3.3†
Alzheimer's	0.2 (60)	-	1.6***	0.1 (433)	1.3‡
*Anxiety**	22.1 (6,782)	31.4 (4,766)	31.2****	1.6 (6,945)	5.9†
Schizophrenia	0.3 (78)	-	.7-1.5	0.1 (573)	0.2-0.59§
Bipolar Disorder	2.9 (886)	-	4.4****	0.2 (1,064)	2.0†
Cardiovascular Disease					
Atrial fibrillation	9.5 (2,919)	-	2-9	3.6 (14,839)	1.2-1.3
Coronary heart disease	14.3 (4,396)	-	6.0	5.0 (20,539)	3-4
Myocardial infarction	5.5 (1,702)	1.1 (161)	4.7 **	3.0 (12,099)	.87-2.46
Obesity	33.7 (10,351)	37.3 (5,662)	39.8	2.6 (10,820)	26.2
Diabetes	21.4 (6,571)	4.8 (724)	12.6	5.0 (20,260)	6.2
Cancer					
Colorectal	2.6 (806)	0.1 (17)	4.2****	1.1 (4,627)	5.3-7.1 ****
Breast (female)	12.4 (2,025)	1.1 (115)	12.4****	5.7 (12,680)	12.5 ****
Lung	2.3 (707)	0.1 (9)	6.2****	0.5 (2,243)	5.9-7.7 ****
Pancreatic	1.0 (313)	-	1.6****	0.2 (749)	1.4 ****
Melanoma of skin	6.2 (1,896)	-	2.3****	0.7 (2,724)	1.9 ****
Prostate (male)	12.4 (1,794)	0.4 (17)	11.2****	3.6 (6,762)	12.5 ****
Bladder	3.7 (1,147)	-	2.3****	0.6 (2,433)	0.9-2.6 ****
Non-Hodgkins lymphoma	3.1 (937)	-	2.1****	0.4 (1,827)	1.7-2.1 ****

∇ Phenotypes were defined using ICD-based PheWAS codes[139] for MGI and UKB and based on survey responses for GFG. A description of the phenotype definitions can be found in **Supplementary Section S2**.

* Any anxiety disorder; ** adults 40 and older; *** adults 65 and older; **** lifetime risk of developing disease/condition; † past week prevalence, refers to the presence of symptoms in the past week; ‡ point prevalence, refers to the prevalence measured at a particular point in time (proportion of persons with a particular disease at a point in time); § estimate is from England

Notes: ranges for schizophrenia represent the minimum and maximum point estimates from several estimates included in the source material; ranges for myocardial infarction and cancer estimates provided indicate the range of sex-specific point estimates

Sources for US and UK estimates can be found in **Supplementary Section S6, Table S3**.