Supplementary Materials to

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

Authors: Lauren J Beesley, Maxwell Salvatore, Lars G. Fritsche, Anita Pandit, Arvind Rao, Chad Brummett, Cristen J. Willer, Lynda D. Lisabeth, Bhramar Mukherjee

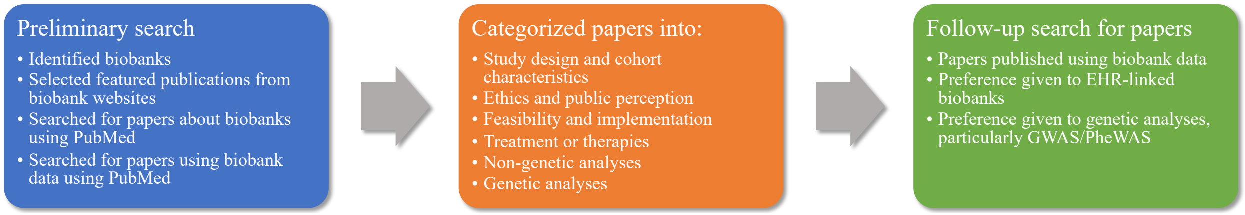
Section S1. Literature search: about biobanks and biobanking and biobank-based research

In this section, we describe the methods used to identify and classify recent literature based on major biobanks. A preliminary search of biobanks was conducted as described in **Figure S1**. First, a university-sponsored database was searched for papers on biobanks and papers published using biobank data. Second, we compiled a short list of biobanks and searched their websites for biobank-promoted research articles. These papers were read to identify various topics for search terms. We identified the following topic areas: study design and cohort characteristics, ethics and public perception, feasibility and implementation, treatment or therapies, genetic analyses (e.g. GWAS/PheWAS), and non-genetic of biobank data.

PubMed was the primary database used. We searched for various combinations of terms related to the topic areas we identified as well as the names of specific biobanks. Papers promoted on biobank websites were also included. Papers from these searches were included if they (a) analyzed data from a biobank (genetic or non-genetic), (b) were published about a specific biobank, or (c) were published about biobanks in general. Papers were excluded if they were not in English, but we placed no restrictions on date of publication (while there was a preference for more recent publications) or geographic region. A subsequent search was conducted focusing solely on papers published using biobank data (particularly biobanks linked with EHR) and performing a genetic analysis. Preference was given to studies where genotype data was analyzed (largely GWAS/PheWAS). The publication search was concluded June 1, 2018.

We would like to emphasize that this is not intended to be an exhaustive list of all biobank-related literature. It was, however, intended to provide a good understanding of the state of biobank literature in general.

Figure S1: Paper Identification Algorithm



Preliminary search
- Identified biobanks
- Selected featured publications from biobank websites
- Searched for papers about biobanks using PubMed
- Searched for papers using biobank data using PubMed

Categorized papers into:
- Study design and cohort characteristics
- Ethics and public perception
- Feasibility and implementation
- Treatment or therapies
- Non-genetic analyses
- Genetic analyses

Follow-up search for papers
- Papers published using biobank data
- Preference given to EHR-linked biobanks
- Preference given to genetic analyses, particularly GWAS/PheWAS

Recent Biobank-Based Literature

We reiterate that this is not intended to be an exhaustive list of biobank-based literature. Through our search, we identified three goals of papers published about biobanks and biobanking: (1) study design and cohort characteristics, (2) ethics and public perception, and (3) feasibility and implementation. Papers analyzing biobank data were grouped in three more categories: (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,[1–3] cohort characteristics,[3–8] how the cohort differs for the rest of the country's population,[9] and characteristics of specific patient populations (e.g. clinical characteristics of colorectal[10] and prostate[11] cancer patients in the BioBank Japan cohort). This information is critical for determining generalizability of results obtained using biobank data. Moreover, these manuscripts are of particular interest to statisticians as they describe the scope of the data and highlight key considerations that have implications for data analysis and interpretation.

Ethics and Public Perception

Attention has been given to ethics of biobanks, particularly to ethical and legal concerns[12] with the use of broad consent (seeking consent for future unspecified research) and to the use of opt-out consents in biobanks with plans for broad, long-term use.[13,14] Additionally, research has looked at public perception of biobanks and biobanking,[15] 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[16] and China[17] that there is general public support for biobanks and large-scale cohort studies.

Feasibility and Implementation

Literature about biobanks explores feasibility and implementation for establishing biobanks, including business plans and models for facilitating biobank creation,[18] how to recruit and obtain consent (particularly among particular groups of patients such as cancer patients),[19–21] and the use of electronic consent in biobanking.[22] Increasingly, biobanks are augmenting their survey data with EHR data. The promise and utility of EHR data for secondary research use has been well-established.[23,24] Research into EHR data quality suggests a need for standardized methods of EHR data quality assessment[25] and awareness of underlying data collection processes.[26]

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.[27] EHR-linked biobank data, particularly those linked with prescription claims data, are well positioned to explore treatment or therapy related 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.[5,28–50] We group these papers published using biobank data into two coarse categories: *genetic* and *non-genetic analyses*. Examples of non-genetic analyses include Song et al. (2018), where the authors describe the protective nature of alcohol consumption on coronary artery disease risk in the Million Veterans Program, and Peters et al. (2018), where they highlight sex differences in the association between measures of general and central adiposity and risk of myocardial infarction in the UKB.[39,44] Pilling et al. (2017) is an example of a genetic study, where the authors conducted a GWAS of UKB data to identify 25 loci associated with human longevity.[51] In a recent paper, Nielsen et al. (2018) used biobank data to explore the relationship between genetics and atrial fibrillation.[52]

Figure S2 provides a distribution of included biobank-based publications falling into each of the above categories over time. The rise in the number of genetic studies can be partly explained by the increase in the number of GWAS and 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, which may depend on the kinds of data available and their willingness to share data externally. **Table S1** details 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, which 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. A discussion regarding some of the more common health-related outcomes studied using biobank data can be found in **Supplementary Section S2**.

Figure S2. Overall Distribution of Selected Biobank-Based Publications by Year and Type

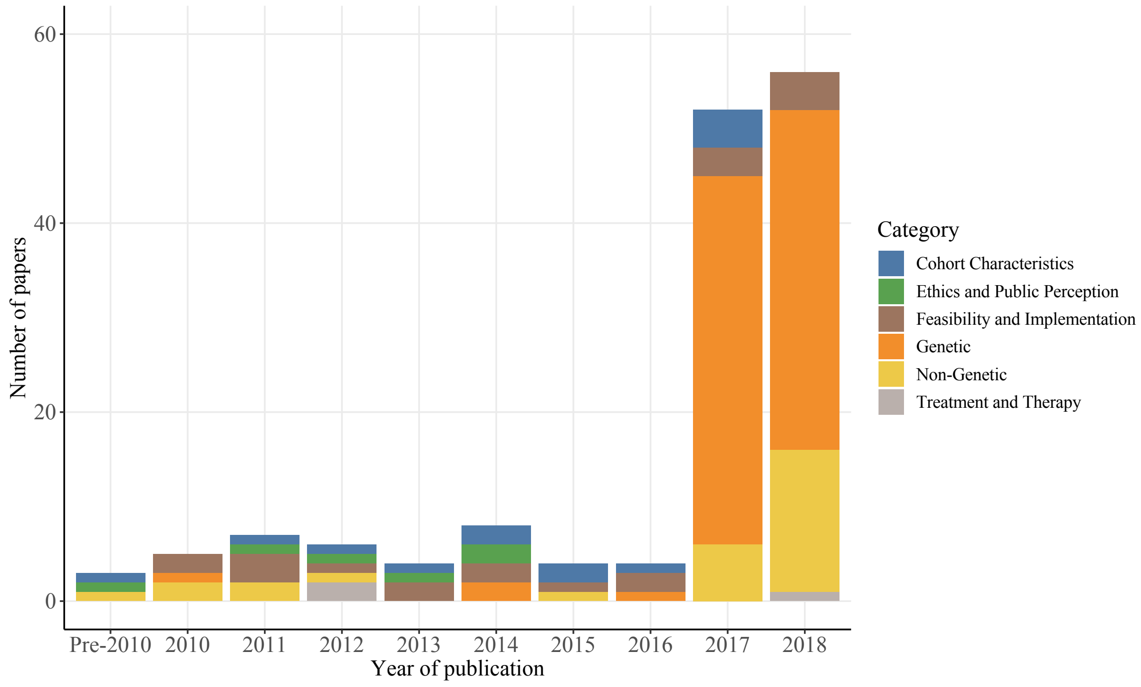


Table S1: Identified Publications by Major Biobanks Included in this Paper

Biobank	**# in review**	**%**	**Pre-2014**	**2014**	**2015**	**2016**	**2017**	**2018**
UK Biobank	58	39%	0	0	2	1	24	31
BioVU	8	5%	6	0	0	0	1	1
BioBank Japan	6	4%	0	0	0	0	6	0
Guangzhou Biobank Cohort Study	6	4%	5	1	0	0	0	0
HUNT	3	2%	2	0	0	0	0	1
China Kadoorie Biobank	3	2%	1	0	0	0	0	2
Michigan Genomics Initiative	2	1%	0	0	0	0	0	2
Million Veterans Program	1	1%	0	0	0	0	0	1
Other biobanks	15	10%	1	2	1	1	6	4
Meta-analysis combining multiple biobanks	24	16%	0	1	0	0	13	10
About biobanks/biobanking	23	15%	10	4	1	2	2	4
TOTAL	**149**	**100%**	**25**	**8**	**4**	**4**	**52**	**56**

Note: This table contains papers identified using the described literature search methods, but it is *not* intended to be an exhaustive list of publications from each biobank. Papers were assigned to a biobank if that biobank's data was used in the paper's primary analysis.

Section S2. 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, depression/major depressive disorder, Alzheimer’s disease, anxiety, schizophrenia, and bipolar disorder. 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, which are defined as a combination of more specific conditions including myocardial infarction. Related conditions like stroke, atrial fibrillation and calcific aortic valve stenosis have also been explored in the literature. Obesity (and related measurements like BMI and waist-to-hip ratio) and diabetes have also been explored. Colorectal, breast, lung, pancreatic, and skin cancers as well as pulmonary conditions including smoking and airflow obstruction have been investigated, but to a lesser extent. Abbreviated citations of papers reporting the outcomes discussed here are presented in **Table S2**.

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.

Table S2. Common Outcomes in Biobank Research

Outcome	Papers
Cancer	
Breast	Anderson et al. 2018; Abana et al. 2017; Hoffman et al. 2017
Colorectal	Morris, Bradbury, Cross, Gunter & Murphy 2018; Usher-Smith et al. 2018
Lung	Hatlen, Grønberg, Langhammer, Carlsen & Amundersen 2011
Pancreatic	van Duijnhoven et al. 2018
Skin	Fritsche et al. 2018
Cardiovascular disease	
Atrial fibrillation	Li et al. 2018; Nielsen et al. 2018; Tikkanen et al. 2018; Ritchie et al. 2010
Calcific aortic valve stenosis	Thériault et al. 2018
*Coronary artery disease/ coronary heart disease**	Deary et al. 2018; Peters, Bots & Woodward 2018; Song et al. 2018; van der Harst & Verweij 2018; Wood et al. 2018; Chan et al. 2017; Klarin et al. 2017; Liu, Erlich, & Pickrell 2017; Lyall et al. 2017; Nelson et al. 2017; Warren et al. 2017; Hagenaars et al. 2016; Jiang et al. 2010
Stroke	Malik et al. 2018; Rutten-Jacobs et al. 2018; Lee et al. 2017
Diabetes/Obesity	Astley et al. 2018; Beaumont et al. 2018; Gill et al. 2018; Peters, Bots & Woodward 2018; Turcot et al. 2018; van Zon et al. 2018; Zegnini et al. 2018; Emdin et al. 2017; Klarin et al. 2017; Liu, Erlich, & Pickrell 2017; Lyall et al. 2017; Márquez-Luna, Loh & Price 2017; Paré, Mao, & Deng 2017; Rask-Andersen et al. 2017; Tachmazidou et al. 2017; Tyrrell et al. 2017; Zhao et al. 2017; Cronin et al. 2014; Arora et al. 2011
Psychiatric/Neurologic	
Alzheimer's diseases	Deary et al. 2018; Gibson et al. 2017; Lin et al. 2017; Smeland et al. 2017; Hagenaars et al. 2016
Anxiety	Strawbridge et al. 2018; Du Rietz et al. 2017; Ward et al. 2017
Bipolar disorder	Deary et al. 2018; McElroy et al. 2018; Strawbridge et al. 2018; Clarke et al. 2017; Croarkin et al. 2017; Reus et al. 2017; Ward et al. 2017; Hagenaars et al. 2016
Depression	Deary et al. 2018; Hall et al. 2018; Howard et al. 2018; Rutten-Jacobs et al. 2018; Strawbridge et al. 2018; Gibson et al. 2017; Howard et al. 2017; Reus et al. 2017; Ward et al. 2017; Wigmore et al. 2017; Hagenaars et al. 2016
Risk-taking behavior	Strawbridge et al. 2018; Du Rietz et al. 2017
Schizophrenia	Deary et al. 2018; Strawbridge et al. 2018; Reus et al. 2017; Ward et al. 2017; Hagenaars et al. 2016
Pulmonary-related outcomes	
Airflow obstruction	Amaral, Strachan, Burney & Jarvis 2017; Wain et al. 2017; Lam et al. 2010
Smoking	Taylor et al. 2018; Amaral, Strachan, Burney & Jarvis 2017; Bjørngaard et al. 2017; Jiang et al. 2010; Lam et al. 2010

* includes myocardial infarction

NOTES: The papers presented in this table were identified as part of the literature search (process described in **Supplementary Section S1**). We reiterate that this review of the literature is not exhaustive. The same paper may appear multiple times if it reports results on multiple outcomes. The papers are list by year of publication and then alphabetically by first author.

Section S3. Characterization of attributes comparing population-based and medical center/health system-based biobanks.

In **Table S3**, we provide a high-level characterization of some features biobank-based researchers should consider and how those considerations differ between population-based and medical center/health system-based biobanks. As with the use of any dataset, it is important to understand the origin and protocol of the data collected and how that impacts analytical considerations. These observations generally apply, but there may be exceptions.

Table S3. Comparison of population-based and medical center/health-system based biobanks

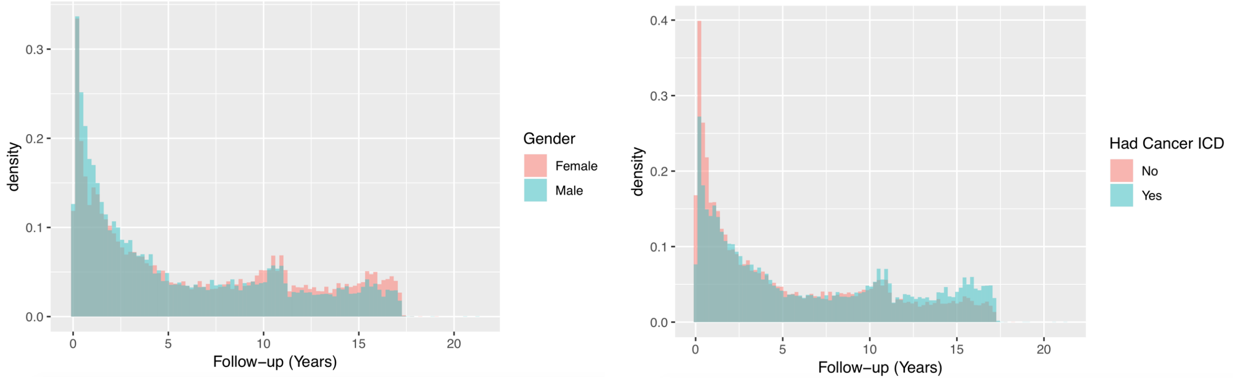
	Population-based	**Medical Center + Health-System Based**
Target population	Representative of population	Varies: geographically restricted to healthcare system catchment area; biased towards sicker individuals
Potential sources of selection bias	Varies; examples include living near an assessment center (UK Biobank) or living in a region of interest (Kadoorie)	Varies; differential ability to overcome access to healthcare (e.g. insured individuals, individuals with access to transportation) and health status (sicker individuals)
EHR: Length of follow-up	Longer due to access to primary healthcare system EHR	Shorter: limited to interactions individual has with medical center/health-system
EHR: heterogeneity	Heterogeneous: complications with variables and different usages and definitions across EHR	Homogeneous: ICD code usage standardized across EHR
Goal	Make inferences about the health of the general population	Make inferences about the health of local region; identify associations with sicker individuals
Examples	UK Biobank, China Kadoorie	MGI, BioBank Japan

Abbreviations: EHR, electronic health record; ICD, international classification of disease; MGI, Michigan Genomics Initiative

Section S4. Description of MGI Patients

In this section, we provide some brief descriptions of the patient populations in used in this study. We note that we restrict our attention to unrelated subjects of recent European ancestry in MGI . We estimated the length of follow-up using the first and last days in which a subject received an ICD code, and the number of visits was defined as the number of unique days in which the subject received at least one phecode. **Figures S3-S5** relate the follow-up time, number of unique phecodes, and number of visits to gender and whether the subject received a cancer ICD code during follow-up.

Figure S3: Follow-up in MGI by Gender and Receipt of Cancer ICD Code During Follow-up



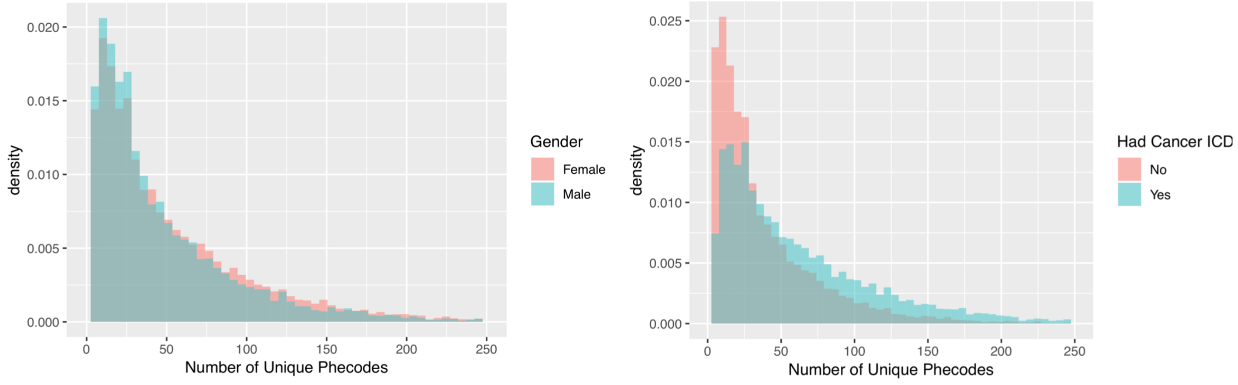
(a) By Gender

(b) By Receipt of Cancer ICD Code

Median (Females): 4.44 years
Median (Males): 3.42 years

Median (No Cancer ICD): 3.31 years
Median (Had Cancer ICD): 4.68 years

Figure S4: Number of Unique Phecodes in MGI by Gender and Receipt of Cancer ICD Code



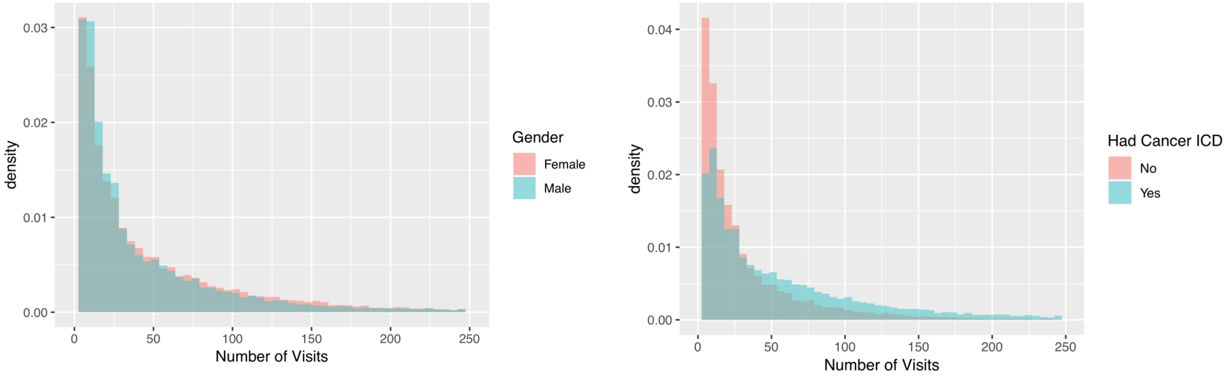
(a) By Gender

(b) By Receipt of Cancer ICD Code

Median (Females): 37 phecodes
Median (Males): 32 phecodes

Median (No Cancer ICD): 25 phecodes
Median (Had Cancer ICD): 46 phecodes

Figure S5: Number of Visits in MGI by Gender and Receipt of Cancer ICD Code



(a) By Gender

(b) By Receipt of Cancer ICD Code

Median (Females): 27 visits
Median (Males): 23 visits

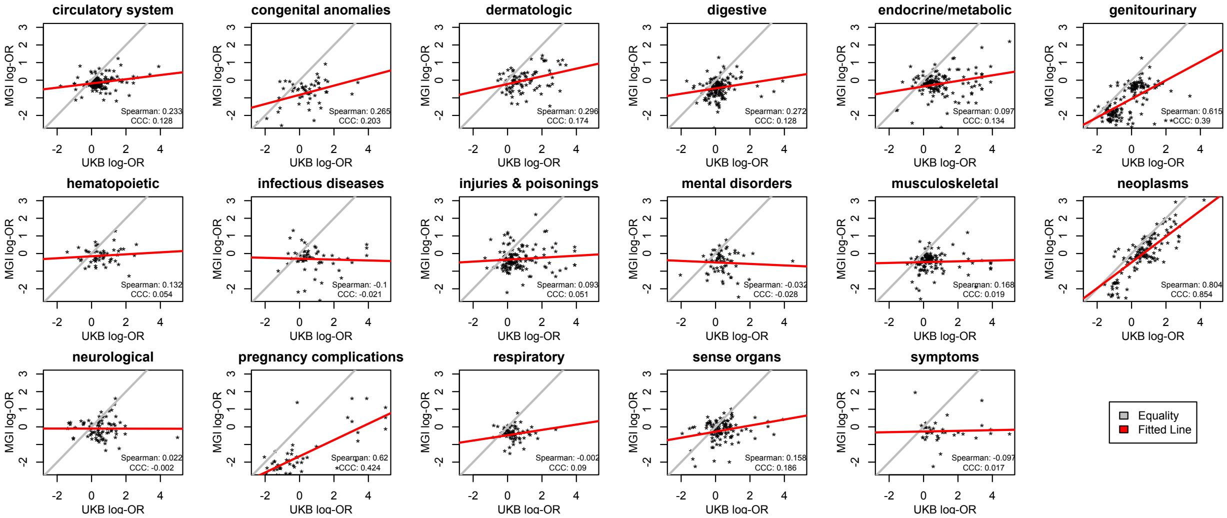
Median (No Cancer ICD): 17 visits
Median (Had Cancer ICD): 38 visits

Section S5. Brief Description of Phenotype Generation

The MGI phenome was based on the Ninth and Tenth Revision of the International Statistical Classification of Diseases (ICD9 and ICD10) code data for 30,702 unrelated, genotyped individuals of recent European ancestry. These ICD9 and ICD10 codes were aggregated to form up to 1,857 PheWAS traits (phecodes) using the PheWAS R package (as described in Fritsche et al. 2018 and Carroll et al. 2014).[55,56] The UK Biobank phenome was based on ICD9 and ICD10 code data of 408,961 genotyped white British individuals that were aggregated to phecodes in a similar fashion as MGI. 1,681 phenotypes (phecodes) were defined in both UKB and MGI.

For each trait and biobank, we identified cases, subjects observed to have that trait. For a given trait, cases were defined as subjects receiving the corresponding phecode at least once during follow-up. Controls were defined as subjects not ever receiving the corresponding phecode. Note that this includes subjects receiving related phecodes. Cases and controls are not matched for this analysis. The prevalence of a particular phenotype (**Figure 2**) was defined as the proportion of subjects receiving a particular phecode in that biobank. In **Figure S6**, the odds ratio of having a particular phenotype (say, Phenotype 1) based on the value of another phenotype (say, Phenotype 2) was computed as $OR = \frac{(n_{11}+0.5)(n_{00}+0.5)}{(n_{10}+0.5)(n_{01}+0.5)}$ using notation in **Figure S7**. The inclusion of the 0.5 terms helps to stabilize odds ratio estimates involving small cell counts.

Figure S6: Log-Odds Ratios of having Melanoma Diagnosis by Other Phenotype Diagnoses*



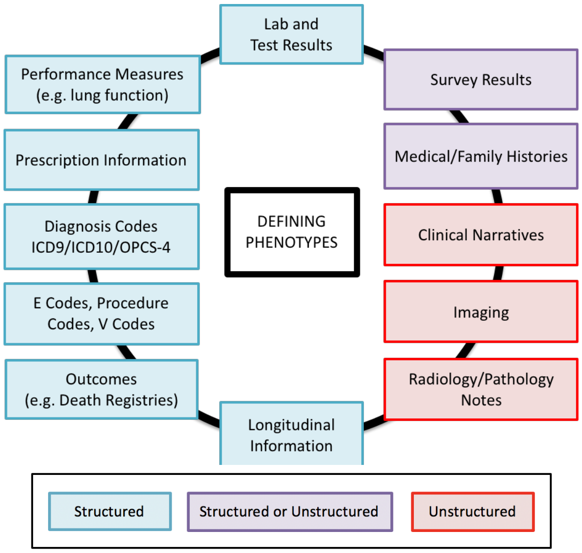
* Each point represents a phenotype in MGI and UK Biobank in a particular disease category (say respiratory) and the corresponding cross-classified log-odds ratios capturing the association between melanoma diagnosis and diagnosis of the other phenotype in MGI and UK Biobank. 1,896 women had observed melanoma in MGI and 2,724 women had melanoma 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 S7: Cross-Tabulation of Phenotypes

	Phenotype 2	
Phenotype 1	No	Yes
No	n_{00}	n_{01}
Yes	n_{10}	n_{11}

While phenotypes in MGI and UKB were generated using ICD codes, future research can consider a broader spectrum of information when defining phenotypes. **Figure S8** provides some examples of additional information in the EHR that may be used to define the phenome.

Figure S8: Potential Data Sources for Generating the Phenome


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

Section S6. Investigating Phecode Definitions and Potential Misclassification

In the process of comparing UKB prevalence estimates to published values for the UK in **Table 2**, we noticed several diseases for which the EHR-derived phenotype codes based on ICD codes in UKB does not appear representative. Most notably, the proportion of subjects receiving ICD codes for obesity in UKB is substantially smaller than the population averages and substantially smaller than the MGI prevalences. In this section, we briefly explore possible causes of this large disparity between EHR-derived phenotypes in UKB and the population averages. We note that the obesity phecode is usually not used in studies with obesity as a primary outcome; rather, researchers usually define obesity using BMI or other measures directly. However, the obesity phenotype may often be used in PheWAS studies considering a large number of phenotypes, and so it is worth exploring potential misclassification of the corresponding ICD-based PheWAS code.

First, we clarify the definitions of the phenotypes. The phenotypes used for the PheWAS and GWAS results, known as phecodes, were derived from ICD codes, but the use of ICD coding varies between MGI and UKB. The available diagnoses of MGI were coded according to the International Classification of Diseases version-9, clinical modification (ICD9-CM) until September 30, 2015 and according to ICD10-CM from October 1, 2015 onwards. All ICD diagnoses were time-stamped, and extracted temporal data were masked as days since birth. Coded ICD values were harmonized to match the formatting used for mapping to PheWAS codes, where trailing characters that are not part of a valid code were trimmed.[56]

The available ICD diagnoses of the UKB were recorded using in-patient hospital admissions, national cancer or death registries. The ICD data was based on WHO's ICD9 codes until roughly 1995 and on ICD10 codes from roughly 1995 onwards, where the ICD9 to ICD10 transition date varied between England, Scotland and Wales as well as between data sources (hospital admissions, death registries, and cancer registries).[57] Where ICD codes contain trailing characters (such as dashes and X's) or other additional characters that are not part of a valid code, UKB applies cleaning rules to strip the trailing characters. Dates of diagnoses were available for cancer diagnoses in cancer registries or for underlying or secondary causes of death (ICD10). "Spell and Episode Data" (admission and discharge) were not readily available for our current phenome-wide explorations.

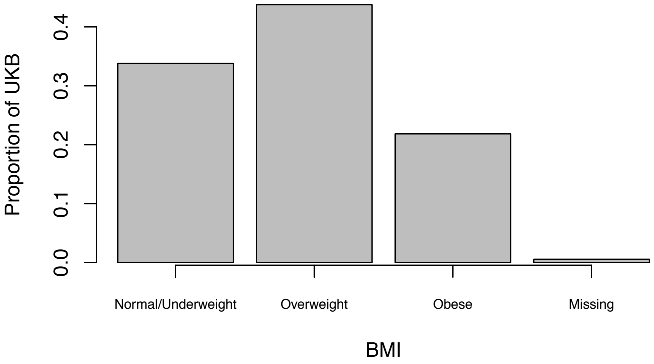
One of the main differences between MGI and the UKB ICD codes is the fact that MGI's diagnoses are based on the ICD9-CM and ICD10-CM coding schemes, which are more extensive than the WHO's original ICD coding schemes.[58] For example, "C44.0" describes the non-melanoma diagnosis "Other and unspecified malignant neoplasm of skin of lip" both in ICD10 and in ICD10-CM. However, there are no ICD10 sub-codes, while the ICD10-CM coding scheme lists the following four sub-codes: "Unspecified malignant neoplasm of skin of lip" (C44.00), "Basal cell carcinoma of skin of lip" (C44.01), "Squamous cell carcinoma of skin of lip" (C44.02), and "Other specified malignant neoplasm of skin of lip" (C44.09). This additional level of detail allows for more granular phenotypes: in this case, the differentiation between basal cell carcinoma and squamous cell carcinoma subtypes. This circumstance is forwarded to the translation of ICD codes to PheWAS codes and is consequently observable in sample size comparisons between MGI and UKB, where PheWAS code subcategories of the latter have markedly fewer or no samples at all (e.g., the PheWAS codes for "Basal cell carcinoma" and "Squamous cell carcinoma" could not be generated from UKB's ICD code data).

Now, we return to the case of obesity. The large difference in the ICD-derived and population proportions of obesity suggest some degree of misclassification of the obesity phenotype based on ICD codes alone. **Figure S9** shows the distribution of (average) BMI values for subjects in UKB. Here, overweight is defined as a BMI between 25 and 30. According to these BMI values alone, we should have at least 21% of subjects being classified as having the obesity phenotype at some point during follow-up. In contrast, only 2.6% of subjects in UK Biobank actually receive ICD codes corresponding to obesity during follow-up.

The MGI phenome does not appear to have such a large gap between the proportion of subjects receiving the obesity phenotype and the expected proportions. One explanation for this phenomenon in UKB is the use of different ICD coding schemes (ICD9 vs ICD10) as described above. For obesity, ICD9 includes codes ("V codes") corresponding to BMI, and these codes are used in the definition of the obesity phenotype. In contrast, ICD10 does not include such BMI-based codes to define obesity. Phenotypes in MGI are often based on ICD9 as many subjects have follow-up prior to implementation of ICD10, while phenotyping in UKB often relies on ICD10, which could partly explain the large differences in observed prevalences between these two biobanks. Additionally, ICD codes related to obesity may be under-reported (so some obese subjects don't get the corresponding code) due to a lack of insurance re-imbursement tied to this code. This misclassification of PheWAS codes could in part explain the disparity between observed and population prevalences for obesity in UKB.

These results provide further motivation for more advanced phenotyping procedures that incorporate additional information outside ICD coding, particularly for diseases in which we believe there will be a large degree of misclassification.

Figure S9: BMI values for subjects in UK Biobank*



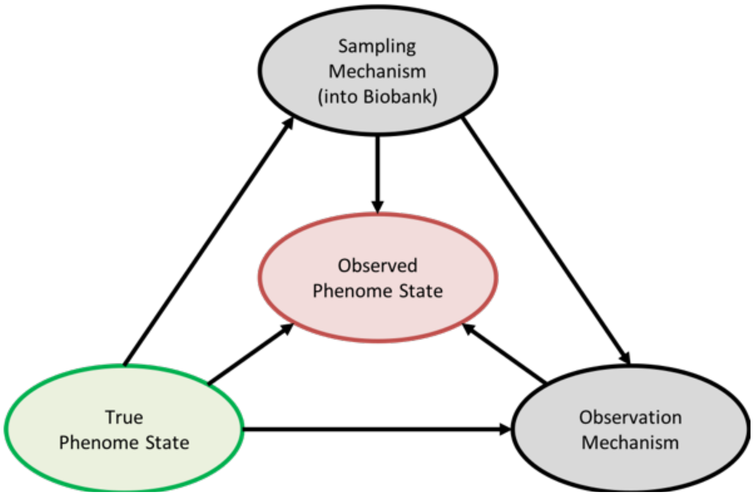
*BMI calculated as the average BMI across 5 visits for which BMI was recorded, listed in UKB data fields 21001 and 23104.

Section S7. Relating Observed Phenotypes to Unobserved Truth

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 S10**, 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.

Figure S10: Relationship between True and Observed Phenome



Section S8. Obtaining and Comparing GWAS Results in the Michigan Genomics Initiative and the UK Biobank

In **Figure 5** of the main paper, we compare GWAS results obtained using MGI and UKB for the "top SNPs" for several different phenotypes. We defined "top SNPs" as described below. GWAS results in MGI and UKB were obtained using the SAIGE method described in Zhou et al. (2018).[59] We considered the following phenotypes: colorectal cancer (phecode 153), prostate cancer (phecode 185), breast cancer (phecode 174.1), and melanoma (phecode 172.1).

For a given phenotype, the "top SNPs" were identified as follows. We first considered all SNPs listed as having reached genome-wide significance for a particular cancer phenotype by the NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas/). We then restricted our focus to SNPs identified by studies in European populations to ensure greater compatibility with the MGI and UKB populations, which are largely of recent European ancestry. No GWAS Catalog studies in the GWAS catalog used MGI data, but some studies may have incorporated UKB data into their analyses.

We then compared GWAS results in MGI and UKB for the subset of SNPs identified by the GWAS catalog with available data in both MGI and UKB. SNPs with minor allele counts less than 3 in either dataset (MGI or UKB) were excluded as were SNPs with differences in the risk allele frequency greater than 0.15 between the two datasets. We further excluded SNPs in linkage disequilibrium, excluding SNPs with R^2 greater than 0.1. This resulted in 25 SNPs for colorectal cancer, 75 SNPs for prostate cancer, 94 SNPs for breast cancer, and 28 SNPs for melanoma. We compare the resulting log-odds ratios from a logistic mixed model fit (from SAIGE) corresponding to the association between a given SNP and the phenotype of interest in a matched subset of the population.

Section S9. Sources for US and UK estimates

Table S4. Sources for US and UK Estimates found in **Table 2**

	US Source	UK Source
Psychiatric/Neurologic		
Depression	National Comorbidity Study	Adult Psychiatric Morbidity Survey
Alzheimer's	Hebert et al. 2003	Alzheimer's Society
*Anxiety**	National Comorbidity Study	Adult Psychiatric Morbidity Survey
Schizophrenia	Jablensky 2000	Kirkbride et al. 2012
Bipolar Disorder	National Comorbidity Study	Adult Psychiatric Morbidity Survey
Cardiovascular Disease		
Atrial fibrillation	CDC	Majeed et al. 2001
Coronary heart disease	CDC MMWR (10/14/2011)	Bhatnagar et al. 2016
Myocardial infarction	Yoon 2016	Bhatnagar et al. 2014
Obesity	CDC	Parliament Briefing 2018
Diabetes	CDC	Diabetes UK
Cancer		
Colorectal	SEER	Cancer Research UK
Breast (female)	SEER	Cancer Research UK
Lung	SEER	Cancer Research UK
Pancreatic	SEER	Cancer Research UK
Melanoma of skin	SEER	Cancer Research UK
Prostate (male)	SEER	Cancer Research UK
Bladder	SEER	Cancer Research UK
Non-Hodgkins lymphoma	SEER	Cancer Research UK

Abbrev: CDC, Centers for Disease Control and Prevention; MGI, Michigan Genomics Initiative; MMWR, Morbidity and Mortality Weekly Report; SEER, Surveillance, Epidemiology and End Results program; UKB, UK Biobank

References

1. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
2. Al Kuwari, H. *et al.* The Qatar Biobank: Background and methods Chronic Disease epidemiology. *BMC Public Health* **15**, 1–9 (2015).
3. 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).
4. Krokstad, S. *et al.* Cohort Profile: The HUNT Study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
5. Jiang, C. Q. *et al.* Smoking cessation and carotid atherosclerosis: the Guangzhou Biobank Cohort Study--CVD. *J. Epidemiol. Community Heal.* **64**, 1004–1009 (2010).
6. 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).
7. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44**, 1172–1180 (2015).
8. Leitsalu, L. *et al.* Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
9. 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).
10. Tamakoshi, A. *et al.* Characteristics and prognosis of Japanese colorectal cancer patients: The BioBank Japan Project. *J. Epidemiol.* **27**, S36–S42 (2017).
11. Ukawa, S. *et al.* Clinical and histopathological characteristics of patients with prostate cancer in the BioBank Japan project. *J. Epidemiol.* **27**, S65–S70 (2017).
12. Greely, H. T. The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks. *Annu. Rev. Genomics Hum. Genet.* **8**, 343–364 (2007).
13. Simon, C. M. *et al.* Active choice but not too active: Public perspectives on biobank consent models. *Genet. Med.* **13**, 821–831 (2011).
14. 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).
15. Ahram, M., Othman, A., Shahrouri, M. & Mustafa, E. Factors influencing public participation in biobanking. *Eur. J. Hum. Genet.* **22**, 445–451 (2014).
16. Starkbaum, J. *et al.* Public Perceptions of Cohort Studies and Biobanks in Germany. *Biopreserv. Biobank.* **12**, 121–130 (2014).
17. Chen, H., Gottweis, H. & Starkbaum, J. Public Perceptions of Biobanks in China: A Focus Group Study. *Biopreserv. Biobank.* **11**, 267–271 (2013).
18. Ciaburri, M., Napolitano, M. & Bravo, E. Business Planning in Biobanking: How to Implement a Tool for Sustainability. *Biopreserv. Biobank.* **15**, 46–56 (2017).
19. Mancini, J. *et al.* Consent for Biobanking: Assessing the Understanding and Views of Cancer Patients. *JNCI J. Natl. Cancer Inst.* **103**, 154–157 (2011).
20. 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).
21. Pillai, U. *et al.* Factors that May Influence the Willingness of Cancer Patients to Consent for Biobanking. *Biopreserv. Biobank.* **12**, 409–414 (2014).

22. Boutin, N. *et al.* Implementation of Electronic Consent at a Biobank: An Opportunity for Precision Medicine Research. *J. Pers. Med.* **6**, 17 (2016).
23. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417–428 (2011).
24. 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).
25. 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).
26. Richesson, R. L., Horvath, M. M. & Rusincovitch, S. A. Clinical Research Informatics and Electronic Health Record Data. *IMIA Yearb.* **9**, 215–223 (2014).
27. 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
28. 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).
29. 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).
30. 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).
31. 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).
32. 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).
33. 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).
34. 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).
35. 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.
36. 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).
37. Kunzmann, A. T. *et al.* Model for Identifying Individuals at Risk for Esophageal Adenocarcinoma. *Clin. Gastroenterol. Hepatol.* **16**, 1229-1236.e4 (2018).
38. 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).

39. 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).
40. 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).
41. 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).
42. 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).
43. Bjørngaard, J. H. *et al.* Heavier smoking increases coffee consumption: findings from a Mendelian randomization analysis. *Int. J. Epidemiol.* **46**, 1958–1967 (2017).
44. 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
45. Ganna, A. & Ingelsson, E. 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study. *Lancet* **386**, 533–540 (2015).
46. 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).
47. 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).
48. Lam, K. H. *et al.* Prior TB, Smoking, and Airflow Obstruction. *Chest* **137**, 593–600 (2010).
49. 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).
50. Yokomichi, H. *et al.* Statin use and all-cause and cancer mortality: BioBank Japan cohort. *J. Epidemiol.* **27**, S84–S91 (2017).
51. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany. NY).* **9**, 2504–2520 (2017).
52. 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).
53. 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).
54. 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
55. 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
56. 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).
57. UK Biobank ICD Coding Information. Available at:

https://biobank.ctsu.ox.ac.uk/crystal/exinfo.cgi?src=Data_providers_and_dates.

58. Jette, N. *et al.* Challenges to the International Comparability of Morbidity Data. *Med. Care* **48**, 1105–1110 (2010).

59. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* (2018).