Supplementary Materials to

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

Section S1. Description of literature search conducted

In this section, we describe the methods we used to identify and classify recent literature based on major biobanks. A preliminary search of biobanks was conducted. 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: feasibility and implementation, ethics and public perception, cohort characteristics, therapy, GWAS/PheWAS, and other analyses 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 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

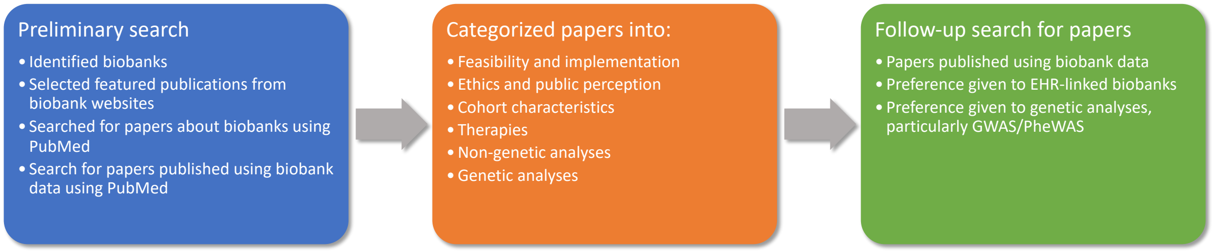

Preliminary search
• Identified biobanks
• Selected featured publications from biobank websites
• Searched for papers about biobanks using PubMed
• Search for papers published using biobank data using PubMed
Categorized papers into:
• Feasibility and implementation
• Ethics and public perception
• Cohort characteristics
• 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

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 biobanks. Papers were assigned to a biobank if that biobank's data was used in the paper's primary analysis.

Section S2. Brief Description of Phenotype Generation

MGI and UKB

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).[1,2] The UK Biobank phenome was based on ICD9 and ICD10 code data of 408,961 genotyped white British individuals that were aggregated to PheWAS traits in a similar fashion as with MGI. A total of 1,681 phenotypes (phecodes) were defined in both UKB and MGI. Additional descriptions of the MGI and UKB phenotyping procedure can be found in **Supplementary Section S5**.

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 3**) was defined as the proportion of subjects receiving a particular phecode in that biobank. In **Figure 4**, 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 S2**. The inclusion of the 0.5 terms helps to stabilize odds ratio estimates involving small cell counts.

Figure S2: Cross-Tabulation of Phenotypes

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

GFG

Phenotyping was done differently for the GFG data. In this biobank, patient phenotype information is self-reported via survey. The depression phenotype was assessed by the following question: “Have you ever been depressed?”. Clearly, the answers to this question may be different than to the question “Have you ever been diagnosed with depression by a physician?”. This explains the comparatively large proportion of patients reporting depression symptoms in GFG compared to MGI and UKB. Anxiety was assessed with the following question: “How anxious are you?” (not at all/slightly/mildly/moderately/severely). We listed a patient as reporting anxiety if they responded with “moderately” or “severely.” Myocardial infarction was assessed with the following question: “Has a physician ever told you that you had a heart attack (a myocardial infarction)?”. Obesity was defined as having a BMI greater than 30, calculated based on self-reported height and weight. Diabetes was assessed with the following question: “Has your physician told you that you had type 1 or type 2 diabetes?”. Cancers were reported by checking boxes indicating cancers the subject has had.

Section S3. Comparison of GWAS Results in the Michigan Genomics Initiative, the UK Biobank, and Genes for Good

MGI and UKB

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).[3] 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.

GFG

We also obtained GWAS results for GFG for the breast cancer phenotype. We were unable to obtain GWAS results for colorectal cancer and prostate cancer due to the small number of cases in GFG (17 cases for each), and the melanoma phenotype was not directly measured in GFG. We further excluded breast cancer SNPs from the X chromosome (not included in the GFG data) and SNPs having a risk allele frequency of less than 0.005 in GFG. This resulted in 86 SNPs used for comparison. We compare the log-odds ratios in MGI with corresponding log-odds ratios in GFG obtained via the SAIGE method in **Figure 6** of the main paper.

Section S4. Description of UKB, MGI, and GFG Patient Populations

In this section, we provide some brief descriptions of the patient populations in MGI, UKB, and GFG used in this study. We note that we restrict our attention to unrelated subjects of recent European ancestry in MGI and GFG and (possibly related) white British subjects in UKB. For 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.

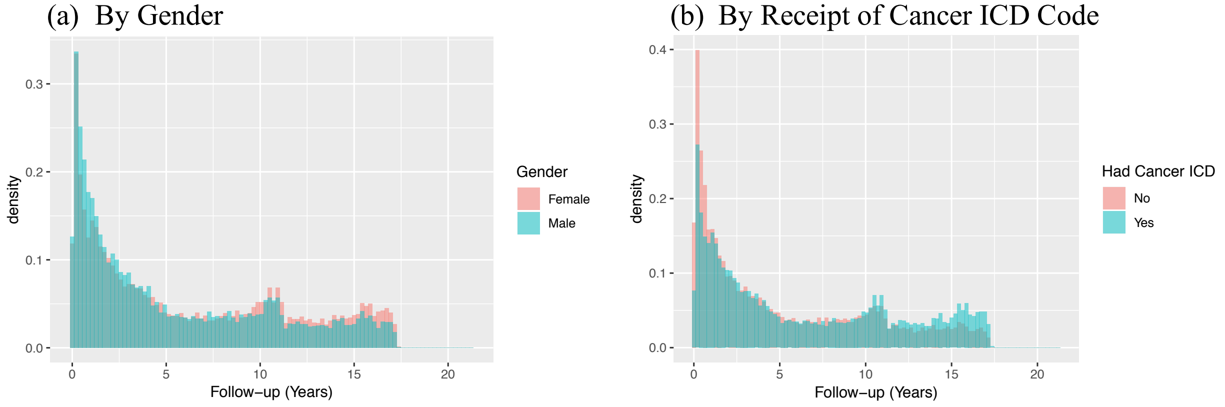
Table S2. Comparison of MGI, UKB, and Genes for Good Patient Populations

	MGI (Academic Medical Center)	UKB (Population-Based)	Genes for Good (Self-Initiated)
Sample Size, n	30,702	408,961	15,156
Females, n (%)	16,297 (53.1)	221,052 (54.1)	10,802 (71.3)
Mean Age, years (sd)	54.2 (15.9)	57.7 (8.1)	36.9 (12.8)
Median Number of Visits Per Participant	27	n/a*	n/a*
Median Days Between First and Last Visit	1,469	n/a*	n/a*
Mean BMI (sd)	29.7 (7.0)	27.4 (4.8)	29.6 (8.1)
Ever Smoked, n (%)	17,044 (55.5)	246,320 (60.2)	10,645 (70.2)**

** Event time data unavailable for UKB and Genes for Good*

** Ever tried a cigarette

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

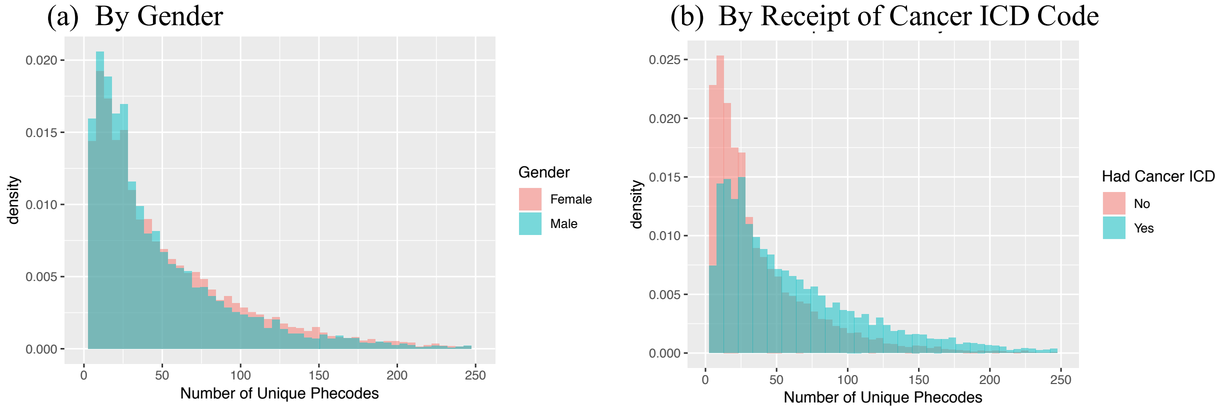

(a) By Gender

(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

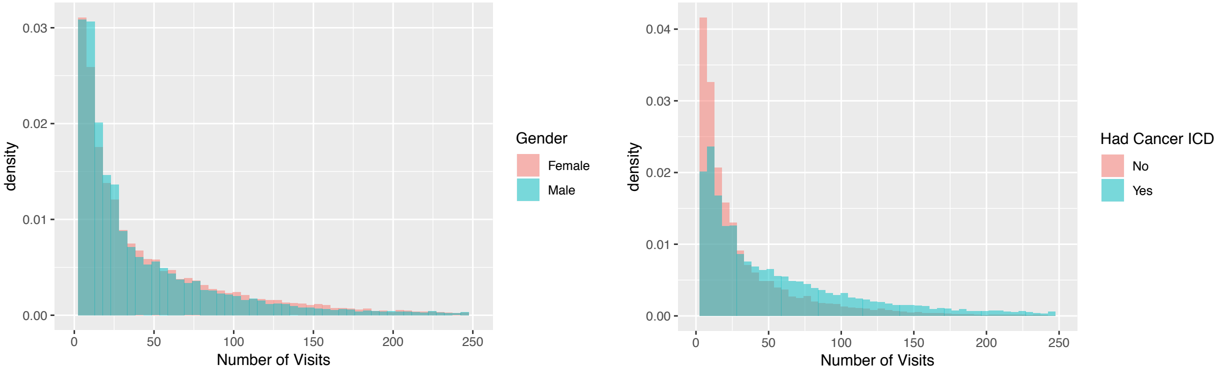

(a) By Gender

(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. 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.[2]

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).[4] 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.

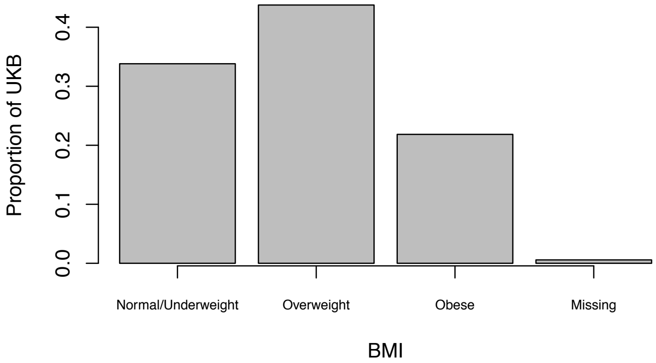
One of the main differences between MGI and the UKB ICD codes is the fact that MGI's diagnoses are based on the ICD9-CM and ICD10-CM coding schemes, which are more extensive than the WHO's original ICD coding schemes.[5] For example, "C44.0" describes the non-melanoma diagnosis "Other and unspecified 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 S6** 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 S6: 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 S6. Sources for US and UK estimates

Table S3: 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. 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
2. 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).
3. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* (2018).
4. UK Biobank ICD Coding Information. Available at: https://biobank.ctsu.ox.ac.uk/crystal/exinfo.cgi?src=Data_providers_and_dates.
5. Jette, N. *et al.* Challenges to the International Comparability of Morbidity Data. *Med. Care* **48,** 1105–1110 (2010).