**Constructing phenotype risk scores for many cancer traits accounting for selection bias in two electronic health record-linked biobanks**

Authors: Maxwell Salvatore, Lauren J Beesley, Ritoban Kundu, Lars G. Fritsche, Christopher R. Friese, Alison M. Mondul, David Hanauer, Celeste Leigh Pearce, Bhramar Mukherjee

**Abstract**

1. **Introduction**

Cancer is a disease of global health concern with a paper recently predicting a 47% growth in cancer incidence globally by 2040.1 While it is characterized by the multiplication of abnormal or damaged cells, a process governed by genes, the causes of cancer cannot be explained by genetics alone. Indeed, a twin study estimates that roughly 30-40% of variance in cancer in the population can be explained by genetics alone,2 meaning most of the variance is explained and related to non-genetic factors. While many worthwhile attempts to summarize genetic contributions of cancer risk (often in the form of polygenic risk scores (PRS)) continue, even if the most powerful cancer PRS can be generated, other factors play a bigger role, emphasizing the need to incorporate other risk factors for personal risk prediction.3

A recent source of data that can be used to aid in risk prediction are electronic health records (EHR). EHR adoption in the US increased through the 2010s and are now commonplace.4,5 They consist of many domains of structured and unstructured data including diagnosis codes, prescription information, laboratory and test results, survey results, medical/family histories, clinical narratives, imaging data, and radiology/pathology notes.6 Structured data, like International Classification of Disease(ICD)-9 and -10 diagnosis codes, are coded data that are already being used to define phenotypes and predict risk. Notably, Denny and colleagues7,8 have aggregated similar ICD-9 and ICD-10 codes into broader yet clinically meaningful phenotypes called *phecodes*. While EHR-linked cohorts and the tools for handling their classification systems (e.g., ICD codes, phecodes), with Big Data comes big responsibility.9 One common challenge to working with EHR-based cohorts is that they are not representative of the target population – an issue known as selection bias.

Building on literature regarding challenges using EHR data,6,10 EHR-based risk prediction,11 and Salvatore and colleagues,12 we use time-restricted phecodes (i.e., phenotypes derived from ICD codes) from EHR at the Michigan Genomics Initiative (MGI; University of Michigan) and UK Biobank (UKB) to predict cancer risk for **X** cancer traits. We compare naïve approaches (i.e., those not accounting for selection) and modern statistical methods (group LASSO, random forest, SuperLearner) with those that use inverse probability weighting and poststratification weighting approaches to account for selection into MGI with the goal of achieving weights (and resulting PheRS) that are generalizable to the broader US population. We contrast the performance of the resulting PheRS, compare PheRS alongside PRS for risk prediction, and discuss the benefits and limitations of each approach. We hope to provide a framework for new researchers interested in using EHR-based risk prediction to think critically about different approaches that can be used alongside other data, like lifestyle factors, laboratory results, and genetics.

1. **Methods**
   1. Data sources

The Michigan Genomics Initiative (MGI) is an electronic health record-linked biobank at University of Michigan Medicine (MM).13 It began in 2012 and recruited adults through surgical and diagnostic pre-/peri-operative visits that required anesthesia. As of the latest data freeze (#4, July 2021), approximately 84,000 participants have consented to (a) access to their electronic health record data (like diagnosis codes [International Classification of Disease or ICD] and laboratory results) for research use, (b) provision of biospecimens for genotyping, (c) data linkage to other datasets like prescription and insurance claims data and neighborhood-level data via their residential address, and (d) recontact for future research use. The Michigan Genomics Initiative and its partner studies (also through Michigan Medicine) recruit approximately 10,000 new participants every year.

The UK Biobank is also an electronic health record-linked biobank in the United Kingdom funded by the UK Department of Health, the Medical Research Council, the Scottish Executive, the Wellcome Trust, and the Northwest Regional Development Agency.14–16 Recruitment took place from 2006 to 2010 and consisted of a nationwide invitation of adults ages 40 to 69 years who lived near one of their assessment centers, ultimately recruiting 502,413 participants. Like the Michigan Genomics Initiative, participants consent to the use of their electronic health record data and biospecimen data including genotype data, ICD9 and ICD10 codes, sex, inferred White British ancestry, kinship estimates down to the third degree, birthyear, genotype array, and precomputed principal components of the genotypes. Additionally, as part of their participation, participants completed an extensive questionnaire on lifestyle and potentially health-related information including sociodemographic and occupational information, early life exposures, psychological state, and family history of illness. Famously, the UK Biobank has made their database accessible to researchers worldwide.

STILL NEED TO DESCRIBE NHANES

* 1. Data
     1. Phenome

When a patient goes to a health care provider for a visit (called an encounter), the provider records clinical diagnoses in their electronic health record. The World Health Organization maintains a classification of these diagnoses in the form of the International Classification of Diseases and Related Health Problems (ICD), though some countries make additional modifications (e.g., ICD-10 Clinical Modification [ICD-10-CM] in the US).17 This classification system undergoes periodic revisions, with the latest 11th edition (i.e., ICD-11) being adopted in 2022.17,18 The Michigan Genomics Initiative and the UK Biobank data currently contain a combination of ICD-9 and ICD-10 codes. These code sets contain tens of thousands of granular codes, some of which are extraordinarily specific (e.g., ICD-10 code Z99.12: Encounter for respirator [ventilator] dependence during power failure). Realizing the utility of this code system, a team from Vanderbilt University aggregated similar ICD codes into what they call “phecodes” that represent clinically meaningful traits, reducing the dimension of the data.7 These phecodes are simply diseases, phenotypic traits, and conditions like intestinal infection (phecode 008), pancreatic cancer (phecode 157), and tobacco use disorder (phecode 318). The ICD-to-phecode translation tables for ICD-9, ICD-10, and ICD-10-CM are publicly available via the PheWAS Catalog (<https://www.phewascatalog.org>; **Table X**).8 The ICD code data from the Michigan Genomics Initiative and the UK Biobank will be distilled into 1,866 phecodes in the latest translation table, which will constitute the medical phenome (or simply phenome). A single occurrence of an ICD code will be sufficient to identify the patient has having the corresponding phecode. The absence of a phecode will be taken to mean the individual does not have the diagnosis. Additionally, the translation table also includes a range of exclusion codes, which disqualify non-cases to be selected as controls for a given trait. For example, for Celiac disease (phecode 557.1), individuals with Celiac disease, ulcerative colitis (phecode 555.2), irritable bowel syndromes (phecode 564.1), and several other related gastrointestinal complaint phecodes (555-564.99) would be ineligible to serve as a control. Data from both cohorts will be time-stamped. For the Michigan Genomics Initiative, all diagnoses are recorded with age (in days) at diagnosis. For the UK Biobank, only the first occurrence of a diagnosis is recorded with age (in days) at diagnosis.

* + 1. Outcome

Cancer diagnoses with at least 500 occurrences in both the Michigan Genomics Initiative and the UK Biobank will be considered as outcomes in these analyses. The cancer phecodes that will be considered are listed in **Table SX**. For example, there were 23 cancer diagnosis phecodes that qualified at this threshold in the Michigan Genomics Initiative phenome from March 18, 2021 (include female breast cancer [n = 5,063], colorectal cancer [n = 2,462], malignant neoplasm of ovary and other uterine adnexa [n = 956], and pancreatic cancer [n = 799]; **Figure SX**). Those with a qualifying cancer phecode will be referred to as cases and the first occurrence of the cancer phecode will serve as the reference age for cases.

* + 1. Time-restricted phenomes

Because the goal is to predict a future cancer diagnosis, diagnostic data needs to be restricted to a time point prior to the occurrence of the outcome. For cases, this is relatively straightforward since the outcome is defined as the age (in days) at the first occurrence of the cancer diagnosis. For reference ages among non-cases I will apply a time-restriction approach first proposed by Salvatore and colleagues.12 Each case will be matched with two controls (i.e., non-cases without an exclusion phecode as described above in SECTION) using Mahalanobis distance matching by age at first EHR diagnosis (nearest neighbor), sex (exact), and length of follow-up (nearest neighbor) to form a matched group. The reference age for each person in a matched group will be that of its case – i.e., the reference age, defined as the age (in days) of the occurrence of the first qualifying phecode, for a case will also be used to restrict the EHR data for both of the case’s controls.

To explore associations and predictive ability, phenome data will be restricted to various time points: 0-, 1-, 2-, and 5- years prior to the reference age. The 0-year threshold will restrict the phenome to diagnoses up to but not including the day of the qualifying diagnosis. While all four thresholds will be considered, the relevant time-threshold for a particular cancer diagnosis will vary. For example, pancreatic cancer prognoses are poor because they are often not diagnosed until after it has metastasized so adequate performance at a lower threshold (e.g., 1-year) could potentially be important. Conversely, because of screening and treatment, breast cancer diagnoses are often at earlier stages with better prognoses where a 5-year threshold may be of more interest.

* 1. Exposures and covariates

In epidemiology, an “exposure” is variable whose association or effect is to be estimated. This term is equivalent to the terms independent variables or predictors in statistics. Here, exposures of interest are any non-outcome diagnoses that could occur prior to the outcome. Specifically, the exposures of interest are the set of all phecodes that (a) occur at least 10 times in both the Michigan Genomics Initiative and the UK Biobank (after time restriction as described in **Section III.d. Time-restricted phenomes**) and (b) are not in the set of exclusion phecodes (as described above in **Section III.a. Phenome**). Covariates that will be included in the analyses include age at restriction (i.e., reference age minus the time threshold), sex, and race/ethnicity.

* 1. Inclusion/Exclusion criteria

Individuals whose age is greater than or equal to 18 years at their last recorded diagnosis will be included. Individuals whose qualifying cancer diagnosis first occurred during their first encounter recorded in their EHR will be excluded from the analysis.

* 1. Analysis

The data preparation described in SECTION and in the following sections (SECTION) will be performed separately for each outcome of interest (cancer diagnosis) and time-restricted phenome. For simplicity, I will use pancreatic cancer (phecode 157) as the outcome of interest through the rest of SECTION though it can be substituted with any of the cancer diagnosis outcomes. SECTION are performed in a discovery cohort while SECTION are performed in a test cohort. First, the Michigan Genomics Initiative will serve as the discovery cohort and with the UK Biobank serving as the test cohort. Then, the roles will reverse.

* + 1. Selection weights

To reduce the effect of selection bias introduced through the primary recruitment mechanism into MGI, inverse probability weights (SECTION) and poststratification weights (SECTION) will be estimated (proposed by Beesley & Mukherjee10). While inverse probability weights are generally preferred, their estimation requires (i) individual-level data, (ii) an adequately specified propensity model, and (iii) sufficient sample size (i.e., can quickly become unstable with too many covariates), poststratification weights are generally easier to estimate because their estimation only requires summary statistics on the population of interest and are non-parametric (i.e., do not rely on proper specification of an underlying model). Selection weights only apply to MGI and not the UK Biobank data.

* + - 1. Inverse probability weighting

Because we are interested in obtaining disease weight estimates that are representative of the US population, we will be using NHANES data as the dataset representing our target population. Because the US population is a large population (i.e., little or no overlap between MGI and NHANES), we can apply a simplified equation (Equation 7 from Beesley & Mukherjee10)

Where is an indicator variable for selection into MGI, is an indicator corresponding to a history of cancer, are additional covariates (e.g., age, sex, race/ethnicity), is an indicator variable for selection into the external sample (i.e., NHANES), and is an indicator variable for selection into the internal (MGI) and external (NHANES) samples (i.e., equal to one for all individuals).

However, because pancreatic cancer is a rare outcome (or may not be available in individual-level data for our external sample), we can implement an alternative approach as proposed by Beesley & Mukherjee that removes from these models and adds an additional term which reflects the ratio of the estimated probability of having in the a history in the internal sample over the estimated probability of having a history of cancer in the external sample such that the equation above becomes

The resulting equation requires the estimation of four models:

* A simplex regression for
* A logistic regression for
* A logistic regression for
* A logistic regression for

For simplicity, for all models, was the set of binary indicator variables: age greater than or equal to 40 and less than 60, age greater than or equal to 60, history of coronary heart disease, history of diabetes, current smoker, former smoker, underweight BMI, overweight BMI, obese BMI, and non-Hispanic White. Additionally, is an indicator representing a history of any cancer, rather than the cancer site of specific interest.

Corresponding selection weights are thus

* + - 1. Poststratification

STILL TO COME

* + 1. Risk score composition

Developing a risk score involves (1) components that comprise the risk score (i.e., variable selection) and (2) estimating the contribution of those components (i.e., variable or “disease” weights). We consider X approaches that can be broken down into two categories of approaches: (1) pruning and thresholding approaches and (2) one-shot approaches.

* + - 1. Pruning and thresholding

These approaches are analogous to pruning and thresholding in the construction of polygenic risk scores [CITE]. First, we performed a phenome-wide, univariable, adjusted *unweighted* screen using a phenome-wide (phecode-phecode) association study (PheWAS). That is, for each of phecodes (exclusive of the outcome, ), we fit the model:

Where is an indicator variable for the outcome, are the set of covariates included in the model (e.g., age, sex, length of follow-up in EHR). We additionally consider a *weighted* PheWAS with the model:

Where the weights are calculated as described above (SECTION).

Next, a p-value thresholding step (the “selection step”) is implemented after the phenome-wide scan is performed. Two multiple testing corrected p-value thresholds will be applied: , where is the number of phecodes considered (and, thus, the number of models fit), and , where is the number of principal components of the phenome data that explain 99% of the variation. In both cases, phecodes whose p-value for is less than the threshold will be selected.

In practice, it is very possible that very few or even no phecodes will be selected at this threshold. We also consider an alternative strategy where phecodes with the 50 smallest p-values are selected as components of the risk score (provided they are also nominally significant at ).

In our naïve approaches, the estimate serves as the disease weight and its p-value is used for pruning and thresholding. The summation of the risk score components and their disease weights are described later (SECTION).

As described thus far, the weights are estimated using a univariable approach. In the *unweighted* PheWAS setting, we consider an additional step where selected phecodes (either by p-value threshold or by smallest p-values) are fit jointly using a weighted Ridge regression. Ridge regression19,20 is a regularized regression method that introduces an L2 penalty and is designed to handle multicollinearity. Ridge regression requires the specification of the parameter, which introduces bias while decreasing variance. The parameter will be selected using cross-validation as the value at which minimizes the mean square error (i.e., optimal performance in the testing data). The ridge-penalized estimates from the joint model will serve as the disease weights.

* + - * 1. Alternative approaches

Increasingly, model statistical methods including regularized regression and machine learning approaches are being used in the development of risk prediction models. Provided that the approach can handle the dimensionality of the data all predictors (in this case, all phecodes and covariates) can be considered and the models will explicitly (e.g., remove or not consider unimportant factors) or implicitly (e.g., downweight unimportant factors) perform disease weight selection and contribution in one step. While these approaches excel at prediction, they (i) are less interpretable, (ii) often do not allow for the incorporation of selection weights, and (iii) are relatively unfamiliar to researchers in epidemiology. Because our goal is prediction, we also consider three additional approaches: group LASSO, random forest, and SuperLearner. These methods do not allow for incorporation of selection weights, so instead we consider selection weights as a possible feature for selection.

Group LASSO, like ridge regression, is a regularized regression approach, but with a different penalty. While ridge regression does not perform variable selection (i.e., does not remove features given to the model), LASSO does. Group LASSO is a modification of LASSO that allows for variable selection in the presence of collinearity within phenotype groups across the phenome. When the phecode-based phenome was curated (built on work by Denny and colleagues7 and defined in the R package PheWAS21), phenotypes were grouped into one of 17 phenotype categories. All phecodes and covariates will be given to the group LASSO model along with each phecodes corresponding category, and the results selected features and their   estimates serve as the disease weights.

Random Forest is an ensemble method based on classification and regression trees. These trees systematically search the predictor space for cutpoints that make classifications (for binary or categorical outcomes) or estimates (for continuous outcomes) that minimize a loss function at each node. Trees have the advantage of being nonparametic and nonlinear and accommodate interactions between variables. Random Forests build many trees using bootstrapped samples of the data (e.g., like Bagging) and randomly selected features to consider at each node.

SuperLearner is an ensemble method that fits many different models (traditional regression, regularized regression, and machine learning approaches, like random forest) and aggregates results across many different models based on their performance.

For these methods, rather than constructing a score as described in SECTION, each models predicted values will serve as the risk score.

* + 1. Scoring

A PheRS will be calculated by summing disease weights (described in SECTION) for phecodes present on an individual’s electronic health record as follows:

where   are the disease weights estimated inSECTION, is an indicator variable for whether person has phecode , and is equal to 1 when the inequality inside is true and 0 when it is false. This is from Salvatore and colleagues,12 where “protective” disease weights are reverse coded, so the resulting risk score is greater than 0.

* + 1. Assessment

Each phenotype risk score will be evaluated for discriminatory ability, calibration, and accuracy using the area under the receiver-operator characteristics curve (AUC), Hosmer-Lemeshow goodness-of-fit test statistic and p-value, and Brier score, respectively. These assessment statistics will be calculated using an unadjusted model. I will also calculate the odds ratio and 95% confidence interval using the 1st, 2nd, 5th, 10th, and 25th percentiles of the phenotype risk score compared to the rest to assess risk stratification ability at the top end of the risk score distribution.

* 1. Software

All analyses will be run using the latest version of R (4.2.0) and packages including SPAtest22,23 (for fast logistic regression in phenome-wide scans), glmnet24,25 (for regularized regression models like ridge), and gglasso26,27 (for group lasso). Scripts will be posted to a public Gitlab repository.

1. **Results**
2. **Discussion**

**References**

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/caac.21660

2. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*. 2016;315(1):68-76. doi:10.1001/jama.2015.17703

3. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19(9):581-590. doi:10.1038/s41576-018-0018-x

4. Charles D, Gabriel M, Searcy T. *Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2014*. Office of the National Coordinator for Health Information Technology; 2015. Accessed January 17, 2023. https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf

5. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital “advanced use” divide. *J Am Med Inform Assoc*. 2017;24(6):1142-1148. doi:10.1093/jamia/ocx080

6. Beesley LJ, Salvatore M, Fritsche LG, et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat Med*. 2020;39(6):773-800. doi:10.1002/sim.8445

7. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma Oxf Engl*. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126

8. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1111. doi:10.1038/nbt.2749

9. Meng XL. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann Appl Stat*. 2018;12(2). doi:10.1214/18-AOAS1161SF

10. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*. Published online December 3, 2020:biom.13400. doi:10.1111/biom.13400

11. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042

12. Salvatore M, Beesley LJ, Fritsche LG, et al. Phenotype risk scores (PheRS) for pancreatic cancer using time-stamped electronic health record data: Discovery and validation in two large biobanks. *J Biomed Inform*. Published online December 2020:103652. doi:10.1016/j.jbi.2020.103652

13. University of Michigan Precision Health. Michigan Genomics Initiative. Accessed February 18, 2022. https://precisionhealth.umich.edu/our-research/michigangenomics/

14. UK Biobank Limited. UK Biobank - UK Biobank. Accessed February 18, 2022. https://www.ukbiobank.ac.uk/

15. UK Biobank Coordinating Centre. *UK Biobank: Protocol for a Large-Scale Prospective Epidemiological Resource*. UK Biobank; 2007. Accessed February 18, 2022. https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf

16. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

17. International Classification of Diseases (ICD). Accessed February 18, 2022. https://www.who.int/standards/classifications/classification-of-diseases

18. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. *BMC Med Inform Decis Mak*. 2021;21(6):206. doi:10.1186/s12911-021-01534-6

19. Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*. 1970;12(1):69-82. doi:10.1080/00401706.1970.10488635

20. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 1970;12(1):55-67. doi:10.1080/00401706.1970.10488634

21. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma Oxf Engl*. 2014;30(16):2375-2376. doi:10.1093/bioinformatics/btu197

22. Dey R, Lee S. SPAtest: Score Test and Meta-Analysis Based on Saddlepoint Approximation. Published online September 4, 2020. Accessed July 3, 2022. https://CRAN.R-project.org/package=SPAtest

23. Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet*. 2017;101(1):37-49. doi:10.1016/j.ajhg.2017.05.014

24. Friedman J, Hastie T, Tibshirani R, et al. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. Published online April 15, 2022. Accessed July 3, 2022. https://CRAN.R-project.org/package=glmnet

25. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1). doi:10.18637/jss.v033.i01

26. Yang (http://www.math.mcgill.ca/yyang/) Y, Zou (http://users.stat.umn.edu/~zouxx019/) H, Bhatnagar (http://sahirbhatnagar.com/) S. gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm. Published online March 18, 2020. Accessed July 3, 2022. https://CRAN.R-project.org/package=gglasso

27. Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalize learning problems. *Stat Comput*. 2015;25(6):1129-1141. doi:10.1007/s11222-014-9498-5