**Constructing phenotype risk scores for many cancer traits accounting for selection bias in an academic medical center-based electronic health record**

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**

**Introduction**

Cancer is a disease of global health concern with a paper recently predicting a 47% growth in cancer incidence globally by 2040.1 It is characterized by the multiplication of abnormal or damaged tissue cells that can spread throughout the body invading and impairing normal tissue and organs, often leading to death. Though the process of cell division is 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 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*.

Building on literature regarding challenges using EHR data,6,9 EHR-based risk prediction,10 and Salvatore and colleagues,11 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 a relatively straightforward regularized regression approach (group LASSO) 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 compare the performance of the resulting PheRS 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.

*Other*…

In the USA, cancer is the second leading cause of death and will remain a leading cause of morbidity and mortality for the foreseeable future.

**Methods**

**Results**

**Discussion**

**References**