Conditional Associations with Big Data: Estimating Adjusted Rank Correlations in the Electronic Health Record

By

Lingjun Fu

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

August 11, 2017

Nashville, Tennessee

Approved:

Bryan E. Shepherd, Ph.D.

Matthew S. Shotwell, Ph.D.

ACKNOWLEDGMENTS

I would like to thank Dr. Bryan Shepherd for the unforgettable experience in the last two years. His probability class is the biased estimate of the best class I have ever taken in my life. This thesis would be impossible without him. He is always glad to help whenever I raise questions in his class or I have obstacles in the thesis work. His attitude towards both life and work sets an excellent example for me.

Special thanks are due to the Dr. Matthew Shotwell for his kind service in my committee. Dr. Jeffrey Blume deserves special mention for granting me the opportunity to join this awesome program. I also appreciate all the graduate faculty in the department for their dedicated work to teaching.

I want to thank all my friends at Vanderbilt for the uncountable moments when we share the happiness and sorrow with each other.

My wife, Xianwen Shen, has always been my source of support, motivation, and backup during this warm and touching journey. Without her, my life would get stuck and I could never become who I am now. I will always remember the old saying: no other success in life can compensate for failure in the home.

Finally, this work is specially dedicated to my parents, who have sacrificed so much to offer me the best opportunities.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

LIST OF TABLES v

LIST OF FIGURES vi

ABSTRACT vii

Chapter

1 Introduction and Background 1

1.1 Introduction to Electronic Health Records (EHR) 1
1.2 Motivation for EHR Studies 2
1.3 Cumulative Probability Models (CPM) 5
1.4 Probability Scale Residual (PSR) 7
1.4.1 Residual 7
1.4.2 Probability Scale Residual 8
1.5 Covariate-adjusted Spearman's Rank Correlation 10
1.5.1 Calculation of Standard Error 13
1.6 Summary 14

2 Application of Covariate-adjusted Rank Correlation to VUMC EHR Data 15

2.1 VUMC EHR data 15
2.2 Data Analysis 16
2.2.1 Data Extraction and Pre-processing 16
2.2.2 Descriptive Statistics 17
2.2.3 Methods 22
2.3 Results: Estimation and Inference 25

3 Conclusions 33

3.1 Summary 33
3.2 Future Research 34

Appendix A Summary of ICD Codes 35

Appendix B Discussion on Conditional Correlation 37

REFERENCES 38

LIST OF TABLES

Table Page

2.1 Criteria for Records to be Kept after Filtering 16

2.2 Summary of Demographic Information (Categorical) 18

2.3 Summary of Demographic Information (Continuous) 18

2.4 Summary of Lab Values (Ordinal) 19

2.5 Summary of Lab Values (Continuous) 20

A.1 Summary of ICD codes 36

LIST OF FIGURES

Figure Page

2.1 The Impact of Binning on Estimation of P-values 24

2.2 Heat map of the pairwise covariate-adjusted Spearman's rank correlations for ICD codes and laboratory tests in the VUMC EHR 28

2.3 Screenshot of the Difference between Adjusted and Unadjusted Spearman's Correlations for Some Labs 29

2.4 Screenshot of the Difference between Partial Rank Correlations for Males and Females for a Subset of ICD Codes 30

2.5 Screenshot of the Difference between Partial Rank Correlations for Males and Females for the 50 Labs 31

2.6 Difference between Partial Rank Correlations for Patients with and without Diabetes 32

ABSTRACT

In this thesis, we apply and adapt a new method to assess conditional associations in a large dataset from the Vanderbilt University Medical Center Electronic Health Record (EHR). We estimate pairwise rank correlations among disease status and lab values in the EHR after adjusting for demographical information. Our covariate-adjusted rank correlations involve fitting cumulative probability models (CPMs), extracting probability-scale residuals (PSRs) from these models, and computing the sample correlation between PSRs for different outcomes. This approach is rank-based, robust, and applicable to a variety of data types. Computational challenges arise with large datasets, particularly when we apply these methods to continuous outcome variables such as most lab values; we propose some workaround solutions. We present our results with estimates and confidence intervals for the partial Spearman's rank correlations among all pairwise combinations of the most frequent 250 ICD codes and 50 lab results among 472,570 patients with data in the EHR. We also present results stratified by sex and diabetes status, demonstrating how to assess for differences in correlations between different population strata.

Chapter 1

Introduction and Background

1.1 Introduction to Electronic Health Records (EHR)

An electronic health record (EHR) is a digital version of a patient's medical chart [1]. EHR are real-time, patient-based records that make information available instantly and securely to authorized users. An EHR system is built to go beyond standard clinical data collected in a provider's office and can be inclusive of a broader view of a patient's care. EHR can:

- contain a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results;
- allow access to evidence-based tools that providers can use to make decisions about a patient's care;
- automate and streamline provider workflow.

An EHR system integrates data for different purposes [2]. It enables the administrator to utilize the data for billing purposes, the physician to analyze patient diagnostics information and treatment effectiveness, the nurse to report adverse conditions, and the researcher to discover new knowledge.

EHRs are much better transformation tools of healthcare information compared with paper-based systems, and they can benefit the whole healthcare system in several ways. From the financial perspective, EHR enhance revenue by reducing costs (supply, transcription, utilization of tests). From the clinical perspective, EHR improve productivity, clinician satisfaction, accuracy of diagnosis and care, quality and convenience of care, patient safety, coordination of care, legal and regulatory compliance, reliability, and aggregation of data and interoperability. Despite these advantages in practice, the adoption rate of EHR is still low in United States and faces obstacles such as physicians' resistance, loss of productivity, lack of standards, and privacy and security concerns [3]. Technically, there are also some challenges to using EHR data: incompleteness, errors, uninterpretable data, inconsistency, unstructured text, selection bias, and interoperability. Nevertheless, the EHR creates great opportunities for clinical and translational research.

1.2 Motivation for EHR Studies

Clinical decision making is complicated because the physician must attempt to bridge what has been referred to as an inferential gap between the information at hand in a given case and what patients need to better their health [4]. EHR can narrow this gap. Typically, structured patient data in the EHR system have four features: demographic information, diagnoses denoted by International Classification of Diseases (ICD) codes, medications (treatment using drugs), and laboratory values. EHR typically also include unstructured patient data that mainly consist of free text notes. Researchers are often interested in assessing associations between different variables in the EHR. A scientific question is what are the internal correlations among features in the EHR, and a statistical question is how to quantify them. Small cohort sizes do not allow a comprehensive view of the co-occurrence of diagnoses, laboratory values, and other patient features. The large sizes of EHR and their tremendous amount of structured patient-level data can help us make inference at the population scale to discover new associations that may be of relevance to patient care. For example, a phenome-wide association study (PheWAS), introduced by Vanderbilt BioVU, uses individual SNPs to check for statistical associations with hundreds of disease phenotypes found in the EHR [5]. Given the ready availability of EHR data, they represent a cost-efficient tool for studying and discovering associations.

There are some common approaches to analyze EHR data. One can do a 2 by 2 table analysis using the chi-square test of homogeneity if the interest is on the comorbidity of two diseases [6]. One can cluster patients based on a similarity metric calculated from feature values [7, 8]. One can perform regression or machine learning methods (e.g. supervised: decision tree classifier or support vector machine (SVM), unsupervised: clustering) to predict future diagnoses [9]. One can also create a cohort study by querying for patients associated and not associated with certain features. With PheWAS, a novel method is proposed to scan phenomic data for genetic associations using ICD codes. However, these methods have their own limitations. The EHR data usually come in large dimensions and are hence sparse, which could induce bias given incomplete and heterogeneous features. While most modern machine learning methods are robust to outliers and errors, they tend to overfit as a result of being purely data-driven. Another drawback of machine learning methods is their difficulty of interpretation. Finally, the various data types in EHR also cause problems since some methods are only appropriate for specific types of variables (e.g. linear regression for continuous variables). Lab values may be skewed or highly variable and require proper transformation before performing analyses. Other metrics may have detection limits.

It is desirable to have methods that can assess for associations in big data (many records and variables). In particular, we need robust methods applicable to a wide variety of outcome types and distributions that can be quickly implemented without requiring transformation and extensive data diagnostics. Spearman's rank correlation is good because it is robust to many outcomes and does not require normality or linear relationships, however, it does not adjust for covariates. In recent work, Spearman's rank correlation has been extended to adjust for covariates [10]. This new method is rank-based, robust, and allows covariate adjustment. It involves three main steps:

- fit robust models (e.g. cumulative probability models) to disease status and lab test values in EHR data;
- obtain probability-scale residuals (PSR) from these models;
- calculate Pearson's sample correlation of PSRs to get the adjusted partial rank corre-

lation.

We can also stratify on variables of interest to get conditional partial Spearman's correlations. For example, the partial correlation between two lab tests among males and among females, after controlling for age, race, and BMI.

The covariate-adjusted rank correlation approach has many strengths. Cumulative probability models are rank-based and good for EHR data because they can handle various data types, complicated relationships, outliers, and censored measurements. The PSRs, which can be easily calculated for all kinds of orderable outcome variables (binary, continuous, ordered categorical, count) can accommodate most variables in EHR data on a common bounded scale between -1 and 1. The Pearson's correlation of PSRs is easy to calculate and does not involve a ranking algorithm. Overall, this approach is easy to implement and achieves a good balance between parametric and nonparametric approaches.

The main goal of this work is to apply covariate-adjusted Spearman's partial rank correlation to study pairwise relations among ICD codes and lab test results while adjusting for demographical variables. There are some potential challenges with fitting covariate-adjusted Spearman's rank correlations in practice. Most of these challenges have to do with computational difficulties with scaling up this method to a big data setting. First, the computational difficulty of fitting cumulative probability models increases as the number of unique values in the outcome variables increases. Second, in order to get inference of the correlation, we need to calculate first and second-order derivatives of the log-likelihood function for each record. Evaluating the inverse of the information matrix is also computationally intensive. Third, lab test values could be unstructured and therefore it is hard to order them without natural language processing (NLP) knowledge. Overall, the challenges in this work lie in applying these methods in a big data setting that stretches the computational capabilities of these methods.

Before going into the details of our study, it is helpful to briefly review the various components of the statistical methods we will be employing.

1.3 Cumulative Probability Models (CPM)

Many variables of interest in EHR are ordinal. That is, we can rank the values, but the real distance between levels is unknown. This type of data distinguishes itself from categorical and continuous data. For example, diseases are graded on scales from least severe to most severe. Some lab test results, especially those recorded by human beings, are recorded in the format of given levels rather than numerical values. When the ordinal variables are predictors with small numbers of levels, in many cases we can treat them as categorical values without losing too much information. The situation is more complicated when we have ordinal response outcomes. In this work, we focus on dealing with ordinal outcomes such as disease status and lab test results. Note that binary outcomes are special cases of ordinal outcomes with the number of levels equal to two.

The general approach for the analysis of ordinal data is to fit cumulative probability models (CPM) which can be motivated by supposing there is a latent continuous variable and that the observed ordinal outcomes come from discretizing the underlying latent variable into K ordered groups. Let the cumulative probability under the specific level k be $\gamma_k(x) = P(Y^* \leq k|X = x)$ such that there is a link function connecting γ_k with the linear model of X:

$$g(\gamma_k) = X\beta + \beta_{k0},$$

where X is a vector of predictor variables and β is a vector of coefficients. "$g()$" is the link function which is often a logit function ($g(x) = \log(\frac{x}{1-x})$) but could be something else depending on the distribution of outcome variables. A cumulative probability model with the logit link is referred to as a proportional odds model, which is so named because it takes the covariate effects to be constant across the K categories. That is, the only varying coefficient is the intercept β_{k0} where k ranges from 0 to $K-1$. The intercepts must increase as k increases to ensure that the cumulative probabilities increase. When the outcome has 2 categories and the logit link is used, this model reduces to logistic regression.

Although generally used for ordinal or binary data, CPM can be fit to continuous data as well [11], which allows us to study most of the data types in EHR. CPM applied to continuous outcomes can be thought of as semiparametric transformation models: a linear relationship is assumed between covariates and a latent variable whose distribution is implied by the choice of link function. CPMs assume that the observed continuous outcome arises from an unspecified monotonic transformation of the latent variable. The "unspecified" transformation explains why it is semiparametric while we still assume a parametric distribution on the transformed latent variable [12]. It has been seen that cumulative probabilities models are fairly robust to minor or moderate link function misspecification with moderate sample sizes. CPMs also directly model the CDF from which important statistics (e.g. moments and quantiles) can be easily calculated. Moreover, CPMs can handle complicated data types and mixed distributions. For instance, some lab test results in the EHR data are continuous in a range and discrete beyond that range. CPMs can be applied as long as the values are still ordered.

Regardless of many attractive features, CPM have not been extensively used to study continuous outcomes. One reason might be computation limits, which are now less of a concern thanks to modern computing power and optimized algorithms. The **orm** function in the **rms** package in R fits CPM for continuous or ordinal response variables, and efficiently allow for a large number of intercepts by capitalizing on the information matrix being sparse [13]. This allows us to model lab tests with thousands of unique values. However, because CPMs with continuous data require estimating $K-1$ intercepts for K unique outcomes, computational challenges still arise when applying these methods to big data such as the EHR. One of the major challenges on the applied side is how to optimize the calculation of the inverse matrix in order to make inference when the model has thousands of intercepts or the sample size is large.

1.4 Probability Scale Residual (PSR)

Whenever a statistical model is fitted to explain the data, it is often important to check for systematic departure from the model. This can be done with diagnostic plots that illustrate the relationship between residuals and predictor variables or outcomes. In addition to diagnostics, residuals may also serve as messengers inherited from the model conveying leftover information of the response variables after accounting for covariates in the model. We are often interested in the correlation between two outcome variables X and Y, after adjusting for potential confounder Z. One way to measure this association is to fit models of X on Z and Y on Z and then to assess the correlation between residuals from these models. For linear models, Pearson's partial correlation is defined in this manner. We will employ a similar technique but fit more general and robust models of X on Z and Y on Z (i.e., cumulative probability models), and compute a more general type of residual, the Probability Scale Residual.

1.4.1 Residual

When assessing a model, residuals serve as a measure of the discrepancy between the observed and fitted values for each observation. The degree to which one observation affects the estimated coefficients is a measure of influence. We can also understand residuals from another perspective: the variation of the outcome variable which is not explained by the current working model.

Pierce and Schafer [14] and Cox and Snell [15] provide nice surveys of various definitions for residuals in Generalized Linear Models (GLM). The general goal for defining residuals is to make their distribution follow some patterns if the model is appropriately specified. This goal is achieved by considering linear equations, link functions, and deviance contributions. In this section, we briefly mention several forms of common residuals.

Response residuals are simply the difference between the observed and fitted values as used in linear regression, which can be treated as "normal scale residuals". Pearson residuals are scaled versions of response residuals. Large absolute values of Pearson residuals indicate failure to fit an observation. The support range of Pearson residuals are not scaled (still the same as Response residuals). Deviance residuals are monotone to the deviance for each observation and based on the χ^2 distribution. In generalized linear modeling, they are generally preferred over Pearson residuals for model checking as their distributional properties are closer to the residuals in linear regression. Score residuals are related to the score function which is the optimized estimating equation.

It is desirable to have a residual that is well defined, easily computable, and robust across many outcome types with a common scale. The residuals listed above have their own advantages and disadvantages in different scenarios depending on the type and distribution of covariates and outcomes. In particular, they are not easily extendable to ordinal outcomes where we lack a natural definition of difference unless we convert the ordinal values to continuous values. Moreover, none of these residuals are suitable for cumulative probability models. These facts motivate the introduction of the PSR.

1.4.2 Probability Scale Residual

As we mentioned earlier, residuals quantify the discrepancy between the observed and fitted values. Alternatively, one can also extend the concept of a residual to a metric quantifying the discrepancy between the observed values and fitted distributions. For example, one can contrast them through their mean, $E(y - Y^*) = y - \hat{y}$, where Y^* is the random variable from the fitted distribution. This is the response residual in linear regression. An extension of the response residual is the mean of the sign of this contrast: $E\{\text{sign}(y, Y^*)\}$, which we refer to as the PSR.

To formally define the PSR, we assume Y is an orderable (binary, count, ordinal, or continuous) random variable from a distribution F. Let y be the observed value of this random variable and F^* be the fitted distribution of Y. The PSR of this observation is:

$$r(y,F^*) = E\{\text{sign}(y,Y^*)\} = pr(Y^* < y) - pr(Y^* > y) = F^*(y^-) + F^*(y) - 1, \tag{1.1}$$

where Y^* is a random variable with distribution F^* and $F^*(y^-) = \lim_{t\to y} F^*(t)$. The support of the random variable $\text{sign}(y,Y^*)$ is $\{-1,0,1\}$. So, the PSR is bounded between -1 and 1. Also, it is monotonic in y and F^* with respect to stochastic ordering. To calculate the PSR, we only need the fitted cumulative density functions (CDF) evaluated at y and y^-, rather than the full specification of the fitted distribution. Note that PSRs are readily calculated from CPMs because CPMs estimate the CDF.

Li and Shepherd [16, 17] introduced the PSR for ordinal outcomes and demonstrated that it has several desirable properties. We emphasize a few of them here. First, the PSR is useful for ordinal outcomes as it does not require assigning arbitrary numbers to categories. Second, it does not require calculation of $E(Y^*)$. Note that although the PSR is calculated using a probability scale, it does not require fully specifying the fitted distribution [18], but rather just cumulative probabilities at some specific values. This property makes it useful with nonparametric and semiparametric models. Third, PSRs are easily calculated for both continuous and ordinal outcomes such that we can compare residuals from different outcomes and models on the same scale (-1 to 1). This property allows us to use PSRs across various features in EHR data. Finally, the PSR has a nice connection with ranks and is easy to interpret.

For continuous outcomes, $r(y,F^*) = 2F^*(y) - 1$. When the model is correctly specified, that is $F^* = F$, the PSR follows a uniform continuous distribution: $r(y,F^*) \sim \text{Unif}(-1,1)$. Correspondingly, the PSR has mean 0 and variance $\frac{1}{3}$. One can use these properties to do model diagnostics. In our EHR data, most lab test results are continuous, although some have detection limits.

For discrete outcomes, $r(y,F^*) = 2F^*(y) - f^*(y) - 1$, where f^* is the probability mass function (PMF) of the fitted distribution F^*. It is proved in [18] that the PSR for discrete outcomes can be viewed as an integrated version of the PSR for some underlying latent continuous variable. In general, however, the PSR in the discrete case does not follow a uniform distribution if $F^* = F$. Its variance, which is equal to $\{1 - \sum f^*(y)^3\}/3$ if $F^* = F$, does converge to 1/3 as the number of discrete levels increases. In our EHR data, all ICD codes (yes or no) and a few lab tests (ordered levels) have discrete outcomes.

For censored outcomes, the PSR is also a function of the censored indicator δ: $r(y, F^*, \delta) = F^*(y) - \delta[1 - F^*(y^-)]$. When we use a fully parametric model to fit the survival data, the calculation of the PSR is straightforward. When we use a semiparametric model (e.g. Cox model), the PSR can be estimated using empirical baseline hazard functions and relative hazards conditional covariates. It is interesting that the three commonly used residuals (Cox-Snell, deviance, and martingale) in the censored case can be written as one-to-one functions of the PSR. Although censored data are found in the EHR (e.g. censored lab values because of detection limits), our analyses do not include censored outcomes.

1.5 Covariate-adjusted Spearman's Rank Correlation

A correlation coefficient measures the extent to which two variables tend to change together. It ranges from -1 to 1, with the sign and value indicating the direction and strength of the relationship, respectively. Pearson's correlation is a parametric linear correlation for two random variables X and Y:

$$\rho_P = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

It is equal to -1 or 1 when and only when there is an exact linear relationship between X and Y: $Y = aX + b$. Spearman's rank correlation is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables can be described using an unspecified monotonic function. For a sample of size n, the n raw values X_i, Y_i are converted to ranks R_{Xi} and R_{Yi} and Spearman's correlation is:

$$\rho_S = \frac{\text{cov}(R_{Xi}, R_{Yi})}{\sigma_{R_{Xi}}\sigma_{R_{Yi}}}.$$

Spearman's correlation is equal to -1 or 1 when and only when Y monotonically increases or decreases as X increases. In the case that $Y = aX + b$, Pearson's correlation and Spearman's correlation are equal. When dealing with ordinal outcomes, nonlinear relationships, skewed distributions, and extreme values, Spearman's correlation is generally preferred.

It is often of interest to assess pairwise associations after adjusting for the influence of other variables. For example, demographic factors such as patients' ages and sex may have large impact on the disease diagnosis and lab test values in EHR data. It may be important to control for these when assessing associations. Pearson's partial correlation is defined as the correlation between response residuals from linear models of X on Z and Y on Z. It can also be written as:

$$\rho_{XY\cdot Z} = \frac{(\rho_{XY} - \rho_{XZ}\rho_{YZ})}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}}.$$

An ad hoc procedure to estimate partial Spearman's correlation is to replace all the Pearson's correlation terms in the above formula with rank correlations [19]. However, this approach does not correspond to a sensible population parameter [20]. We adopt new estimators for covariate-adjusted Spearman's rank correlation proposed in [10]. There are two general estimators: partial and conditional. The partial correlation is the correlation after removing the effect of covariates. The conditional correlation is a correlation written as a function of adjusted covariates. For example, a partial correlation between biomarkers X and Y may be defined as the correlations after removing the influence of other covariates (e.g. sex, age). The conditional correlation may be defined as the correlation between biomarkers X and Y among men and the correlation among women, which may differ. In general, conditional correlations may have more than two levels. In some cases, the marginal partial correlation may be washed out by conditional partial correlations with opposite signs across levels of covariates.

It has been shown that Spearman's correlations, both partial and conditional, can be computed as the correlation of PSRs [10]. In the unadjusted case, that is, in the absence of covariates, the sample correlation of PSRs reduces to the common Spearman's rank correlation of raw values. The PSR of an observed value is fully determined by the empirical CDF, which is based on the ranks of the observed data. Therefore, PSRs are a linear transformation of the ranks. It can be proved that:

$$\gamma_{XY} = \mathrm{corr}[r(X,F), r(Y,G)],$$

where γ_{XY} is the Spearman's rank correlation between random variables X and Y and r denotes the PSR [16].

In the adjusted case where the PSR is calculated from models including all adjusted covariates, the sample correlation of PSRs is equal to Spearman's partial correlation. Specifically,

$$\gamma_{XY\cdot Z} = \mathrm{corr}[r(X,F_{X|Z}), r(Y,G_{Y|Z})].$$

Here, Z represents the covariates to be included to fit model F for outcome variable X and model G for outcome variable Y. The impacts of Z on X and Y are removed through models F and G, from which the PSRs can be calculated and reflect the covariate-adjusted relationship between X and Y. In practice, we model F and G using cumulative probability models.

After adjusting for demographical variables, additional interest may lie in the conditional correlations. For example, the association among diseases and lab tests may be different between males and females. The conditional partial Spearman's rank correlation is defined in [17] as:

$$\gamma_{XY\cdot Z|Z_1} = \mathrm{corr}[r(X,F_{X|Z}), r(Y,G_{Y|Z})|Z_1],$$

where Z_1 is a subset of covariates in Z to be conditioned on. Therefore, it is possible to use PSRs to estimate conditional associations as well.

1.5.1 Calculation of Standard Error

The point estimator of covariate-adjusted partial Spearman's rank correlation is obtained by calculating the Pearson's correlation of two sets of PSRs, each of which comes from a model fitting the outcome of interest on the adjusted covariates. PSRs are functions of the fitted semiparametric distribution and readily calculated from the CPMs we choose.

After obtaining the point estimator, we need to estimate the standard error and p-value. One simple way to do this is to treat the PSRs as random variables and the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

will follow a t distribution with degree of freedom $n-2$, where r is the sample Pearson's correlation and n is the number of pairwise complete observations [21]. This is problematic as the two sets of PSRs are residuals from models with the same set of covariates and hence may be correlated. Recent work [17] proposes two approaches: a bootstrap method and a large sample approximation. In our case, we go with large sample approximation using M-estimation [22]. In short, let $X_{i,res}$ be the PSR for subject i from the model of X on Z and $Y_{i,res}$ be defined similarly. Let $\theta = (\theta_x, \theta_y, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, with $\theta_1 = E(Y_{i,res})$, $\theta_2 = E(X_{i,res})$, $\theta_3 = E(Y_{i,res}X_{i,res})$, $\theta_4 = E(Y_{i,res}^2)$, $\theta_5 = E(X_{i,res}^2)$, and θ_x and θ_y denoting the parameters in the models of X on Z and Y on Z. The goal here is to estimate $\hat{\gamma}_{XY\cdot Z} = (\hat{\theta}_3 - \hat{\theta}_1\hat{\theta}_2)/\sqrt{(\hat{\theta}_4 - \hat{\theta}_1^2)(\hat{\theta}_5 - \hat{\theta}_2^2)}$, whose point estimate can be calculated from the sample Pearson's correlation and variance can be obtained from the delta-method. The large sample distribution of $\hat{\theta}$ under usual conditions is: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[0, V(\theta)]$, where $V(\theta) = A(\theta)^{-1}B(\theta)[A(\theta)^{-1}]'$, $A(\theta) = E[-\frac{\partial \Psi_i(\theta)}{\partial \theta}]$, and $B(\theta) = E[\Psi_i(\theta)\Psi_i(\theta)']$. Note that $\Psi(\theta)$ is composed of many estimating functions (e.g. score equations and moment-based estimating functions). In practice, estimating the large sample distribution of the Fisher's transformation of $\hat{\gamma}_{XY\cdot Z}$, i.e., $\log[(1+\hat{\gamma}_{XY\cdot Z})/(1-\hat{\gamma}_{XY\cdot Z})]/2$, usually leads to more rapid convergence to normality, and is therefore adopted by us to construct confidence intervals.

For conditional correlation, it is easy to calculate the sample Pearson's correlation of PSRs within each level of the categorical variable Z (e.g. gender, diabetes) given that we have sufficient observations in each category. Similarly, standard errors and confidence intervals can be also obtained through M-estimation techniques. P-values can also be computed using M-estimation techniques to test whether conditional correlations across discrete strata are the same.

1.6 Summary

In this chapter we have introduced electronic health record data and covariate-adjusted Spearman's rank correlations using CPMs and PSRs. In Chapter 2, we will apply the described statistical methods to examine pairwise partial rank correlations between a large number of variables in the VUMC EHR. Although these methods have been applied in other settings, they have never been applied to such large sample sizes and numbers of variables. These induce computational challenges which we highlight and provide workarounds for in our EHR analysis.

Chapter 2

Application of Covariate-adjusted Rank Correlation to VUMC EHR Data

Now we dive into the main analysis of this work, which studies the correlation among various variables in EHR data after adjusting for demographical factors. We first introduce the VUMC EHR, describe our dataset, and finally proceed to calculate partial and conditional partial rank correlations across a wide variety and number of EHR variables. Various challenges and our solutions will be highlighted in our analyses.

2.1 VUMC EHR data

Vanderbilt University Medical Center (VUMC) is known for its highly acclaimed teaching hospital and its groundbreaking efforts in EHR. Over 65 million notes were stored in VUMC EHR system during the period 1999-2008, with 50.1 million stored during 2004-2008 [23]. Scanned documents in the EHR come from a diverse set of sources, including clinical notes (35%), administrative documents (28%), orders and prescriptions (19%), and lab test results (10%).

Among massive information available in the VUMC EHR data, two important pieces of data are ICD codes and lab test results. ICD is the foundation for the identification of health trends and statistics globally, and the international standard for reporting diseases and health conditions [24]. Uses of ICD include monitoring of the incidence and prevalence of diseases, insurance reimbursement, and keeping track of safety and quality guidelines. In our case, the patient is assigned an ICD code given a diagnosis and it may evolve as the doctor gains more and better knowledge in diagnosis. As a result, we can get the longitudinal disease history of each patient as well as the prevalence of a specific disease over the population of patients visiting VUMC. Lab tests are ordered to help with diagnosis and treatment decisions. At least 50% of clinical lab tests ordered whose results are in a positive/negative, ordered categorical, or numerical format are incorporated in certified EHR technology as structured data. Often, the occurrence of lab tests are closely related to the assigned ICD codes.

2.2 Data Analysis

2.2.1 Data Extraction and Pre-processing

The data we analyze throughout the whole work is based on the Synthetic Derivative (SD) database containing clinical information derived from VUMC EHR. The SD is de-identified and altered to the extent that it closely resembles the original record, but that identifying features (e.g. dates, names, etc) have been adjusted or stripped. Therefore, the SD can be used as a stand-alone research source. The data consist of three csv files, corresponding to demographic data, ICD codes, and lab test values.

Table 2.1: Criteria for Records to be Kept after Filtering

Entries	Range
weight (KG)	[50, 130]
height (CM)	[150, 200]
BMI (KG/M^2)	[20, 45]
ICD date	01/01/2000 – 08/31/2016
LAB date	01/01/2000 – 08/31/2016
age (years)	≥ 18

We received demographic data from the SD on 582,243 patients. This dataset was filtered for our analyses. The filtered dataset contains demographic data for 472,570 patients. Patient records were kept only if the patient ID showed up in all three databases. Since the methods we will employ are robust to outliers in the outcome variables but not necessarily the predictor variables, we initially filtered the demographical data by removing those records falling out of reasonable ranges. A summary of filtering criteria is shown in Table 2.1. Age was not directly provided in the data; we took the difference of patients' dates of birth and the time of their earliest entry in the database as their age. Based on age, we filtered the data by only keeping adult records (18 years or older). Race was clustered to 5 categories: Asian, Black, White, others, and unknown.

For each patient ID, the same ICD codes may have been assigned multiple times. Similarly, the same lab test may have been performed more than once. For simplicity, we summarized all the distinct ICD codes per patient as never versus ever occuring. For lab tests, the earliest lab test value attached to each patient ID was used if a particular lab was performed more than once. Certainly this collapsing of data results in a loss of information, but it is reasonable in this proof-of-concept study. There were 48,954 distinct ICD codes and 4,235 distinct lab tests. ICD codes included both ICD-9 and ICD-10 codes; we made no attempt to cluster codes in these analyses. We focus on the 250 most frequent ICD codes and the 50 most frequent lab tests with orderable outcomes. While the ICD code results are simple as they are binary (never versus any), the lab test results can be quite messy. First, not all lab test results are orderable (e.g. ABO blood type). Second, many lab tests have string values and typos in recording (e.g. many urine related labs do not have numerical results). Third, some lab test values are censored due to detection limits (e.g. a range instead of a number is given). After consulting with physicians, we chose to drop outcome variables that are not orderable, we ordered string values based on clinical orderings, we corrected typos as best we could, and we incorporated censored values to their nearest thresholds (e.g. the measurement of bilirubin in urine had detection limits at 2 and 12, and we rounded all "< 2" results to 2 and all "> 12" results to 12). We made no effort to filter lab values except ordering those lab results with ordinal string values and converting those continuous lab results to numerical values.

2.2.2 Descriptive Statistics

Based on the cleaned data, a summary of descriptive statistics of demographical information is shown in Table 2.2 and Table 2.3. 58% of included patients were male, 79% white and 10% black, and median age was 46 years. The median BMI was 27.6, which is typically considered overweight.

Table 2.2: Summary of Demographic Information (Categorical)

Factors	number	percentage
Sex		
male	274,383	58
female	198,187	42
Race		
white	372,166	78.8
black	46,157	9.8
asian	6,905	1.5
other	1533	0.3
unknown	45,809	9.7

Table 2.3: Summary of Demographic Information (Continuous)

Covariates	Mean	Median	0.25 percentile	0.75 percentile
weight (KG)	82.6	80.9	69.0	94.5
height (CM)	170.1	170.2	162.6	177.8
age (years)	46.7	46	33	59
BMI (KG/M^2)	28.5	27.6	24.3	31.8

The 50 lab values used in these analyses are described in Table 2.4 and Table 2.5. Two labs were binary (bilirubin and nitrite), four were ordered categorical (shown in Table 2.4), the remaining 44 labs were continuous. With that said, the number of distinct levels for the continuous labs was highly variable, from 11 (urobilinogen) to 7104 (thyroid stim hormone). Although we selected the 50 most common labs, a large proportion of patients still had missing lab measurements, ranging from 20% (creatinine blood) to 60% (cholesterol blood). The 250 most common ICD codes and the proportion of patients with them listed is given in the Appendix A.1.

Table 2.4: Summary of Lab Values (Ordinal)

Lab names	levels	number of observations	percentage
BILIRUBIN	2	195,238	100
	negative	191,076	97.9
	positive	4,162	2.1
URINE BLOOD	5	194,227	100
	negative	141,026	72.6
	trace	14,641	7.5
	small	14,368	7.4
	moderate	11,981	6.2
	large	12,211	6.3
CHARACTER	4	195,332	100
	clear	160,605	82.2
	slight	7,607	3.9
	cloudy	21,419	11
	turbid	5701	2.9
HEMOLYSIS INDEX	4	296,807	100
	none	282,271	95.1
	slight	10,695	3.6
	moderate	2,757	0.9
	gross	1,084	0.4
KETONE URINE	5	185,255	100
	negative	171,495	92.6
	trace	7,890	4.3
	small	3,112	1.7
	moderate	1,575	0.9
	large	1,183	0.6
NITRITE	2	194,591	100
	negative	187,820	96.5
	positive	6,771	3.5

Table 2.5: Summary of Lab Values (Continuous)

LAB names	n obs.	n of distinct levels	mean	median	S.D.	0.25 percentile	0.75 percentile
CHOLESTEROL BLOOD	185027	566	189.1	186	47	159	215
IMM GRANULOCYTES(ABS)	180843	147	0.03	0.02	0.07	0.01	0.03
IMM GRANULOCYTES %	180888	103	0.3	0.3	0.5	0.2	0.4
LYMPHS(ABS)	188380	1005	2	1.8	6.9	1.4	2.4
LYMPHS %	188445	865	25.7	25.9	11.4	17.8	33.1
BASOPHILS %	191549	68	0.4	0.4	0.3	0.2	0.6
BASO(ABS)	191555	68	0.03	0.03	0.08	0.02	0.04
MONO(ABS)	191588	381	0.6	0.6	0.3	0.4	0.8
NEUTROPHILS %	191638	927	63.4	62.7	12.8	54.9	71.8
MONOCYTES %	191646	356	8	7.7	2.8	6.2	9.4
NEUT(ABS)	191602	2859	5.7	4.6	12.5	3.4	6.7
UROBILINOGEN	195810	11	1.4	2	1.1	0.2	2
PH URINE	199318	19	6.1	6	0.8	5.5	6.5
SPECIFIC GRAVITY UA	191754	67	1	1	2.3	1	1
MEAN PLATELET VOLUME	226409	93	10.5	10.4	1.3	9.8	11
THYROID STIM HORMONE	233128	7104	2.3	1.5	7	1	2.4
PROTEIN TOTAL BLOOD	239635	474	7.2	7.2	0.6	6.8	7.6
NT AUTOMATED ABS	239556	3090	5.6	4.6	3.9	3.3	6.6
RDWSD	261918	633	44.7	43.8	5.2	41.5	46.7
ALBUMIN BLOOD	295285	329	4.1	4.2	0.5	3.9	4.4
ALKALINE PHOSPHATASE BLOOD	311541	997	82.2	72	59.7	58	90
BILIRUBIN TOTAL BLOOD	310489	548	0.7	0.6	1.1	0.4	0.8
ALANINE AMINOTRANSFERASE BLOOD	315047	1159	34.9	23	120.1	16	33
ASPARTATE AMINOTRANSFERASE BLOOD	317531	1203	35.7	23	160.6	19	30
NUCLEATED RBC#	199810	195	0.01	0	0.3	0	0
NUCLEATED RBC	179689	62	0.07	0	2.2	0	0
ANION GAP	362754	81	9.2	9	2.9	7	11
PLATELET COUNT	362645	1006	252.3	245	84.8	202	294
RED CELL DISTRIBUTION WIDTH	368327	247	13.7	13.3	1.8	12.8	14.1
MEAN CORPUSCULAR HEMOGLOBIN CONCENTRATION	368445	185	33.4	33.4	1.3	32.6	34.2
MEAN CORPUSCULAR HEMOGLOBIN	368461	330	29.9	30.1	2.4	28.9	31.2
RED BLOOD CELLS	368502	701	4.6	4.6	2.2	4.2	4.9
MEAN CORPUSCULAR VOLUME	368665	501	89.5	90	5.9	86	93
HEMOGLOBIN	368818	206	13.5	13.7	1.8	12.5	14.8
WHITE BLOOD CELL	368712	2238	8.3	7.4	6.5	5.9	9.5
CALCIUM BLOOD	372938	153	9.3	9.3	0.6	9	9.7
CHLORIDE BLOOD	373601	85	103.6	104	3.7	102	106
CARBON DIOXIDE BLOOD	373811	257	26.3	27	3.2	24	28
UREA NITROGEN BLOOD	373912	189	14.9	13	8.9	10	17
SODIUM BLOOD	373958	76	139.1	139	3	138	141

Continued on next page

Table 2.5: Summary of Lab Values (Continuous) ***(continued)***

LAB names	**n obs.**	**n of distinct levels**	**mean**	**median**	**S.D.**	**0.25 percentile**	**0.75 percentile**
POTASSIUM BLOOD	374275	99	4.1	4.1	0.5	3.8	4.4
PACKED CELL VOLUME BLOOD	378419	396	40.6	41	4.9	38	44
GLUCOSE BLOOD	378644	728	107.7	95	46.7	86	110
CREATININE BLOOD	378830	1463	1	0.9	0.9	0.7	1

End

2.2.3 Methods

A randomly picked small subset was used for initial exploratory data analysis (EDA) and model validation. After this was done, the full filtered dataset was used for analysis. Analyses were run on the clusters at the Advanced Computing Center for Research & Education (ACCRE).

First, we estimated the partial Spearman's rank correlation for all pairwise combinations of the 300 ICD codes/labs. In our case, the adjusted covariates Z included the demographical variables: gender, race, BMI, and age. We first fitted CPMs for each ICD code (binary) and lab test (continuous, binary, and ordinal) with Z. To make the models more flexible, we treated gender and race as categorical variables and fitted cubic splines for BMI and age. Specifically, for each outcome Y, we fitted a CPM with the logit link:

$$\log\frac{P(Y\leq k)}{1-P(Y\leq k)} \sim \mathrm{SEX}+\mathrm{RACE}+\mathrm{rcs}(\mathrm{BMI},3)+\mathrm{rcs}(\mathrm{AGE},3)+\beta_{k0},$$

where the range of k is determined by the rank distributions of the observed outcome variables. For example, ICD code is binary with 0 indicating no and 1 indicating yes. Therefore, $k=1$, and we fitted one intercept in addition to the coefficients for covariates. When the model was applied to continuous variables such as most of the lab test results, we needed to fit the same number of covariate coefficients but more intercepts, with $k+1$ being equal to the number of unique outcome values. This significantly slowed down the computation when the test values had large numbers of unique values (4 out of 50 labs in our data had more than 2000 unique values; specifically, they were neutrophil antibody with 2859 unique values, thyroid stim hormone with 7104 unique values, NT automated ABS with 3090 unique values, and white blood cell with 2238 unique values). We grouped those lab test results with more than 1000 unique values down to 1000 quantiles and conducted simulations to investigate the effect of grouping. Our simulations suggested that correlation point estimates remain similar after reducing the number of unique outcomes from 2000 to 1000 by binning. Certainly the validity of this approximation might be case by case depending on the distribution of lab values, but it is reasonable in the context of 50 labs we studied as the numbers of unique values for most of the labs affected by the binning were only slightly above 1000 (evidenced in Table 2.4).

To apply M-estimation techniques to obtain confidence intervals and p-values, we needed to calculate more expressions in addition to PSRs. Specifically, let n be the number of valid observations for a biomarker and L be the number of distinct levels across all observations. We needed to calculate nL first order derivatives of log likelihood, L^2 second order derivatives of log likelihood, and nL first order derivatives of PSRs and then invert the large information matrix. This added more computational burden. As a workaround, we reduced L by grouping lab test results with more than 200 unique values down to 200 quantiles when computing confidence intervals and p-values. Note that about 50% of the lab tests had fewer than 200 unique values. It is interesting to explore the impact of binning on estimation of p-values. We investigated this by comparing the p-values estimated by grouping to 200 quantiles versus 100 quantiles. Fig. 2.1 shows the comparison of log of p-values in two different binning methods (200 versus 100). Most of them appear to be consistent.

Second, we estimated partial rank correlations conditional on gender and conditional on diabetes status. Both covariates are binary. Diabetes was defined in these analyses as having at least one of the following ICD codes: “250.00”, “250.01”, “250.02”, and “E11.9”. We inherited the model outputs from the calculation of partial correlations. Also, in theory, we needed to fit new models adding diabetes status as a covariate when we evaluated the conditional correlation on diabetes. In practice, the difference in PSRs were small (see discussions in Appendix B), so it was justified to just use the outputs from the CPMs including demographical covariates only.

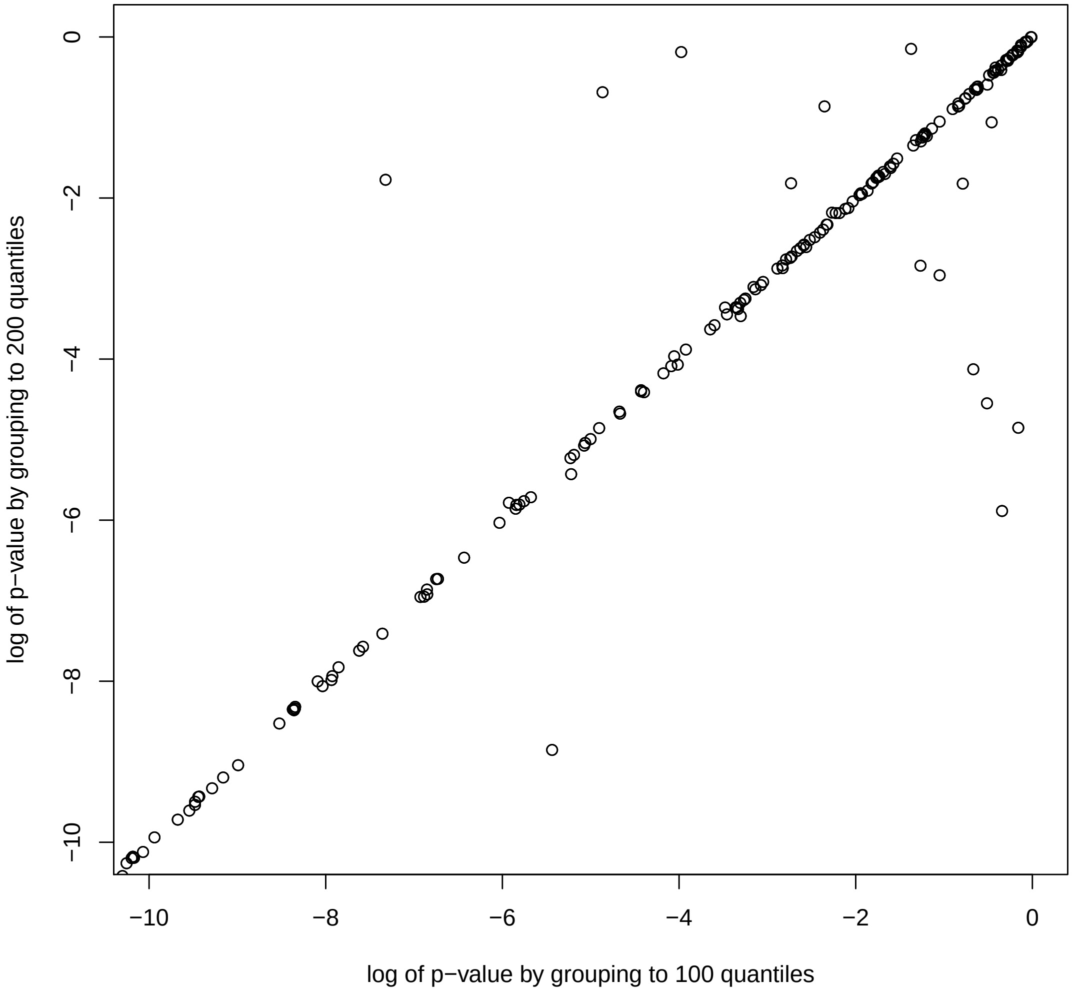

Figure 2.1: The Impact of Binning on Estimation of P-values

2.3 Results: Estimation and Inference

We calculated the point estimates and confidence intervals of partial correlations among all pairwise combinations of the most frequent 250 ICD codes and 50 lab results, and then compared them with the unadjusted Spearman's rank correlation. Any strong contrast between the adjusted and unadjusted correlations may suggest the impact of demographic covariates on the associations of biomarkers. We also calculated the point estimates and confidence intervals of partial correlations conditional on gender or diabetes status as well as the p-value of testing whether the partial correlations across two strata are the same.

Although Fig. 2.2 provides an overall view of the pairwise correlations, it is helpful to be able to look more closely at specific correlations. We developed a Shiny app to help visualize the results. Fig. 2.2 shows a heat map of the pairwise covariate-adjusted Spearman's rank correlation for the 300 variables in the VUMC EHR data. Q1 and Q2 indicate the pairwise biomarkers. The left bottom square is for the pairwise correlation among 250 ICD codes, and the right top square is for 50 lab values. The left top and right bottom regions are for correlations between ICD codes and lab values. It appears that most correlations among ICD codes are positive, which suggests that a patient diagnosed with some disease tends to be diagnosed with other diseases. There are many strong correlations among the lab values as well. It is interesting that most correlations between ICD codes and lab values are weak. Intuitively, disease is a complicated issue which cannot be inferred from only a single lab test.

In addition to showing pairwise Spearman's correlations, the Shiny app also allows us to see the heat maps for unadjusted Spearman's correlation and the difference between the unadjusted and adjusted correlations. In addition, it includes 95% confidence intervals calculated through M-estimation based on large sample approximations and Fisher's transformation. Investigators can zoom into any specific area in the heat map and click the points to see the summary statistic of biomarkers around this region.

Fig. 2.3 shows a screenshot of the difference between adjusted and unadjusted Spearman's correlations zooming in a particular region. For example, the unadjusted Spearman's rank correlation between red blood cells (RBC) and mean Corpuscular Hemoglobin Concentration (MCHC) is 0.21, whereas the adjusted correlation is 0.09 (95% CI: 0.086, 0.093), suggesting that the strength of association between RBC and MCHC is explained in part by demographical variables.

Another aspect to check our results is through gender-specific diseases. For example, screening for malignant neoplasms of cervix (ICD code "V76.2") is for females only, while screening for malignant neoplasms of prostate (ICD code "V76.44") is for males only. It is not applicable to calculate the correlation between them. Therefore, we set the unadjusted correlations to be NA. Otherwise, one may end up with a strong correlation as they are exclusive from each other. However, the correlations of PSRs are close to 0 because we have removed the impact of gender. Note that normally we would not be able to fit the CPM if the ICD code is gender specific (e.g. for female only). However, there were some errors in recording the patient's gender such that a few males (females) showed up for female (male) only ICD code such that the estimation of coefficients in the model was possible. Nevertheless, this is a good example of diluted correlations after adjusting for demographical variables.

We also examined partial correlations conditional on sex and conditional on diabetes status. The main interest lies in whether the pairwise correlations among two variables differ between males and females or between diabetics and non-diabetics. In addition to the Shiny app for partial correlations, we also developed apps to facilitate the visualization of conditional partial correlations for sex and diabetes. For example, we show the partial correlations for males, females, and their difference.

Fig. 2.4 shows the difference between partial rank correlations for males and females for a subset of ICD codes. We see that the associations between a disorder of bone and cartilage (ICD code "733.90") and other diseases is generally stronger for females than males. In contrast, Fig. 2.5 shows the difference in partial correlations between males and females for the 50 lab values. There is little difference between sexes. It is interesting to note that the occurrences of “Simpson’s paradox”: marginal partial correlation, that is either smaller or greater than both conditional partial correlations. For example, the marginal partial correlation between disorders of bursae and tendons in shoulder region (ICD code “726.10”) and nausea with vomiting (ICD code “611.72”) is 0.027, which is smaller than both the conditional partial correlation for males (0.068) and females (0.075). This paradoxical scenario actually occurred among 4.2% of all pairwise associations.

We also show the partial rank correlations for patients with diabetes, with no diabetes, and the difference. Fig. 2.6 shows the difference between partial rank correlations for patients with and without diabetes. It appears that pairwise partial correlations among ICD codes for patients with diabetes are higher than those without diabetes. For example, the partial Spearman’s rank correlation between Long-term (current) use of insulin (ICD code “V58.67”) and Long-term (current) use of aspirin (ICD code “V58.66”) is 0.246 for diabetics and 0.066 for non-diabetics, suggesting that the strength of association between the use of insulin and aspirin is higher for those with diabetes. Another example is to look at the association between glucose and chloride in blood. It is -0.227 for diabetics and 0.005 for non-diabetics.

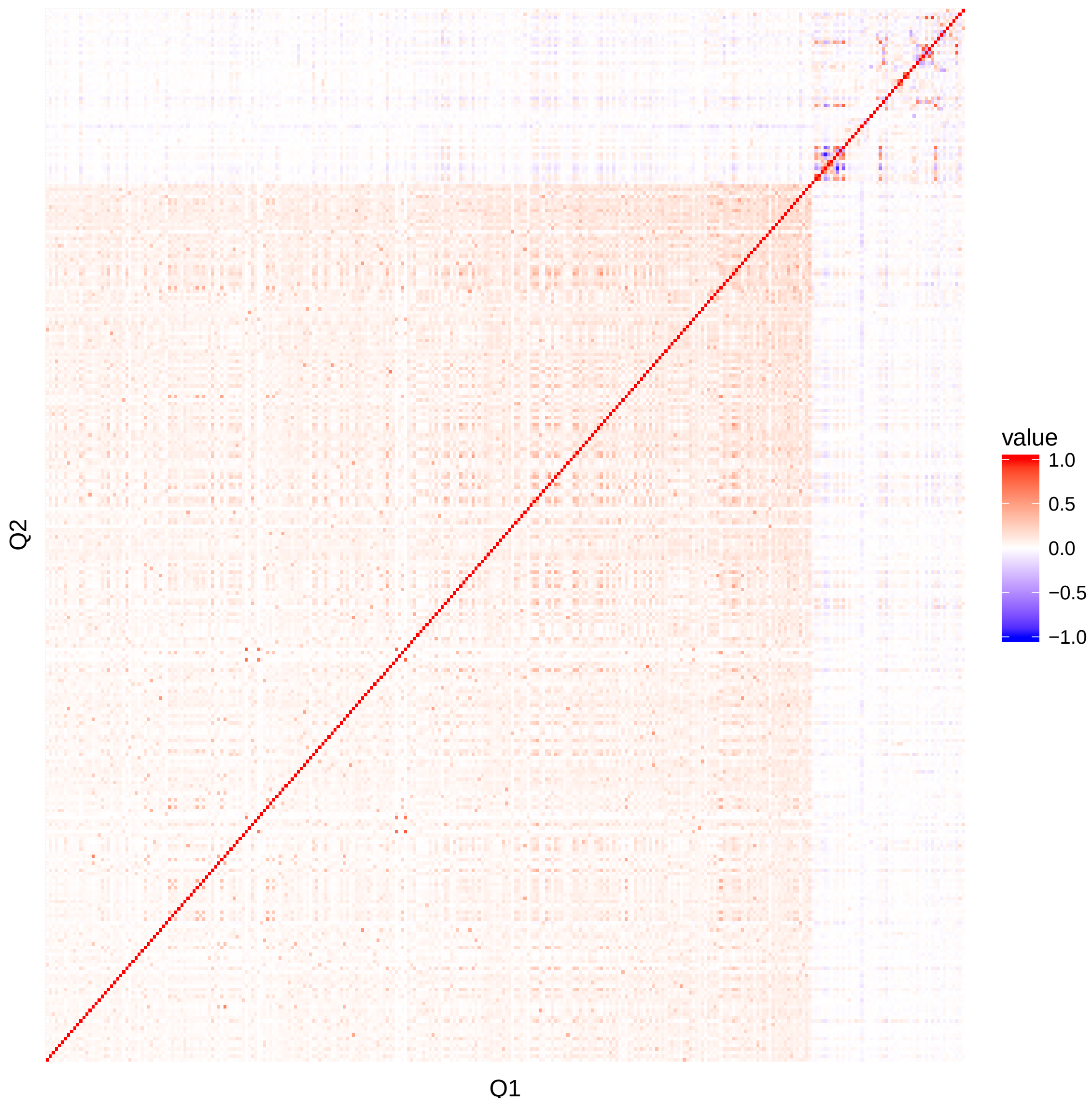

Figure 2.2: Heat map of the pairwise covariate-adjusted Spearman's rank correlations for ICD codes and laboratory tests in the VUMC EHR

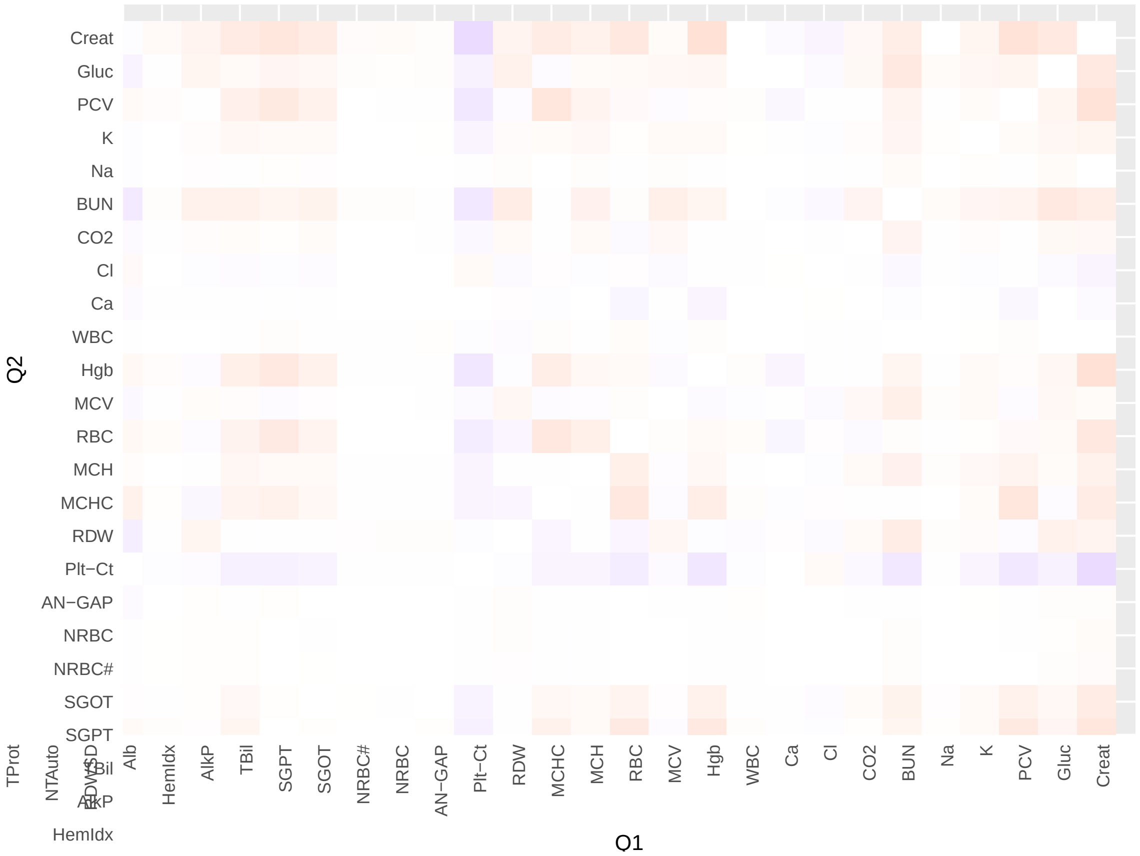

Figure 2.3: Screenshot of the Difference between Adjusted and Unadjusted Spearman's Correlations for Some Labs

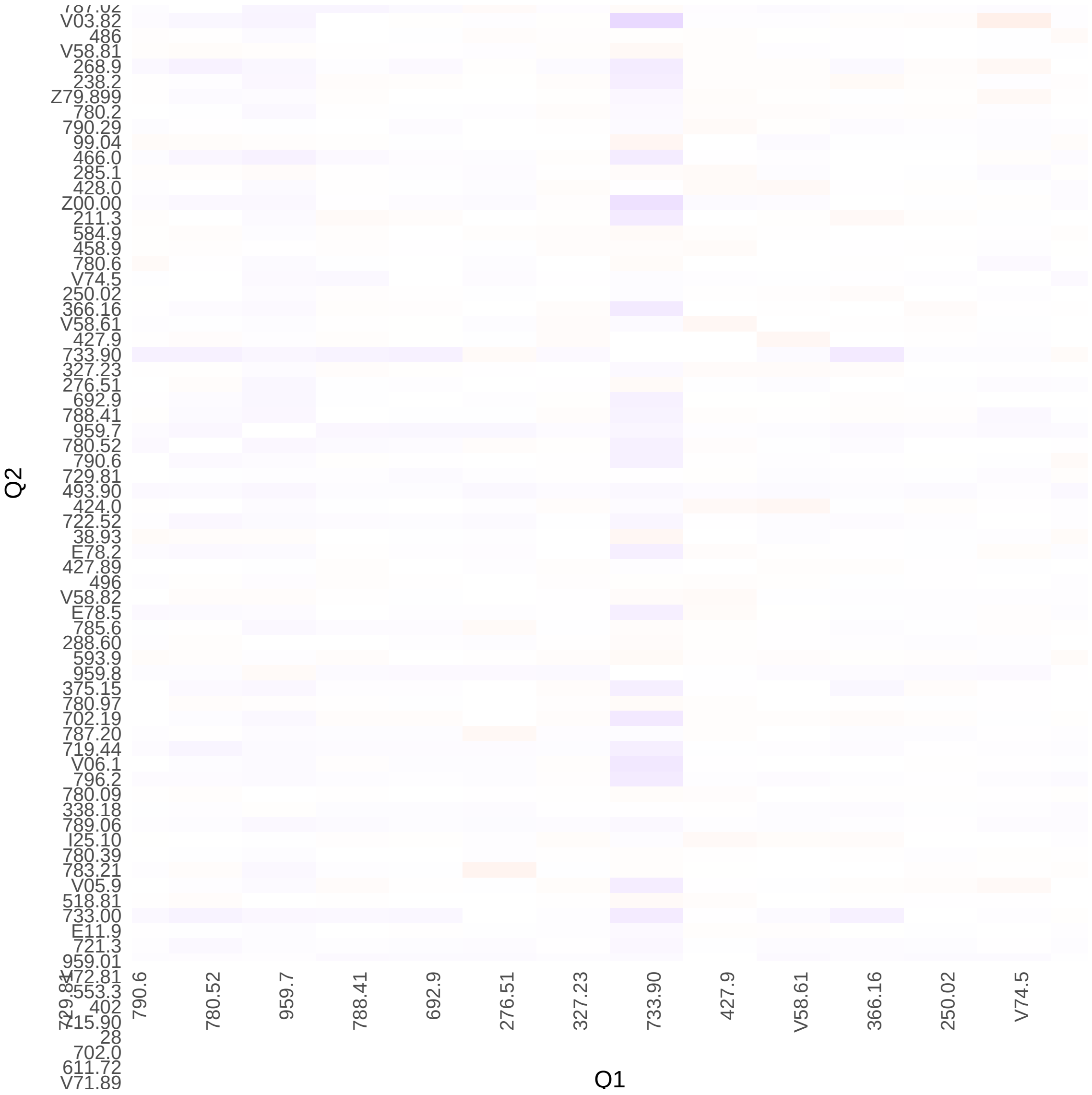

Figure 2.4: Screenshot of the Difference between Partial Rank Correlations for Males and Females for a Subset of ICD Codes

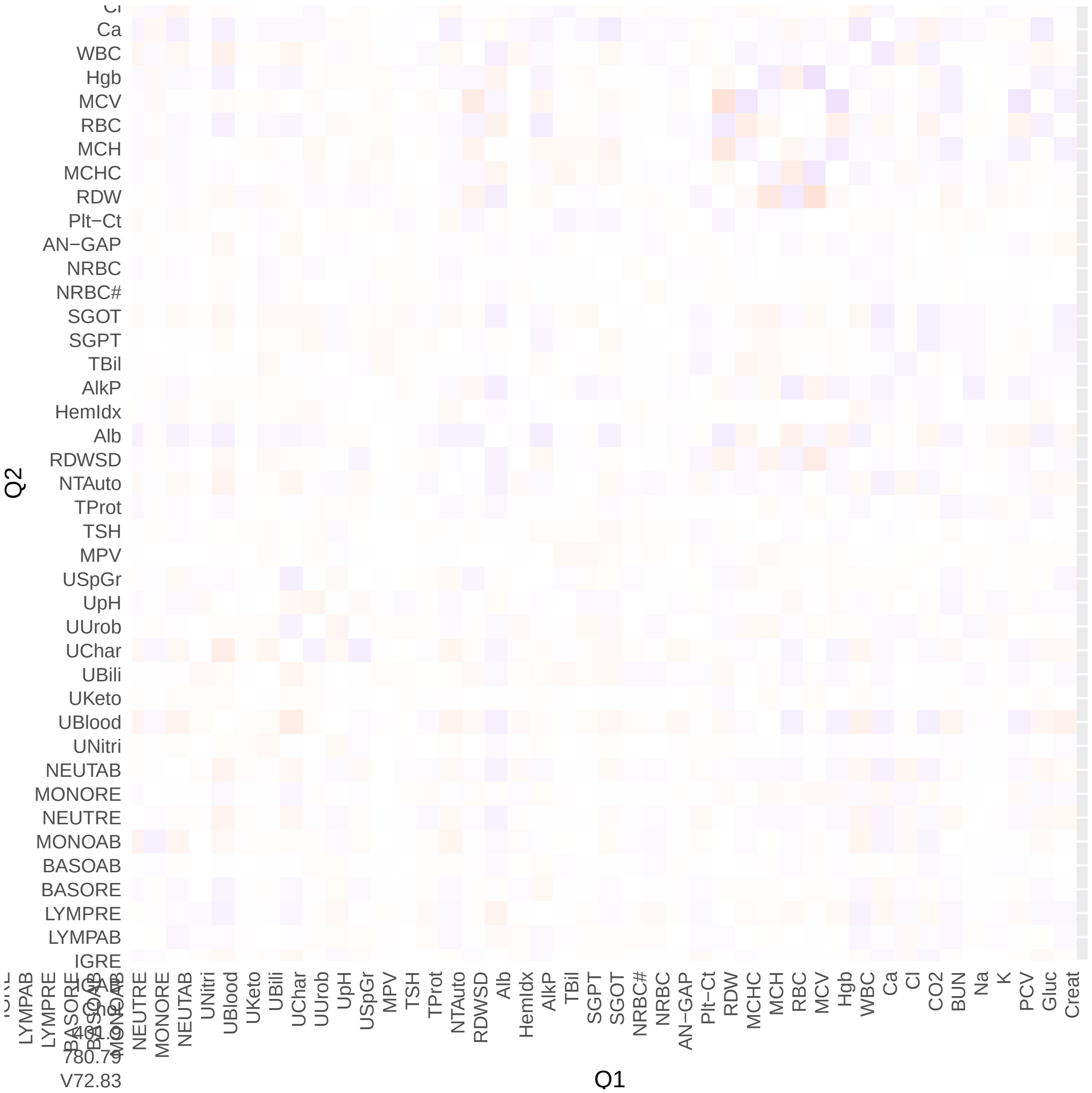

Figure 2.5: Screenshot of the Difference between Partial Rank Correlations for Males and Females for the 50 Labs

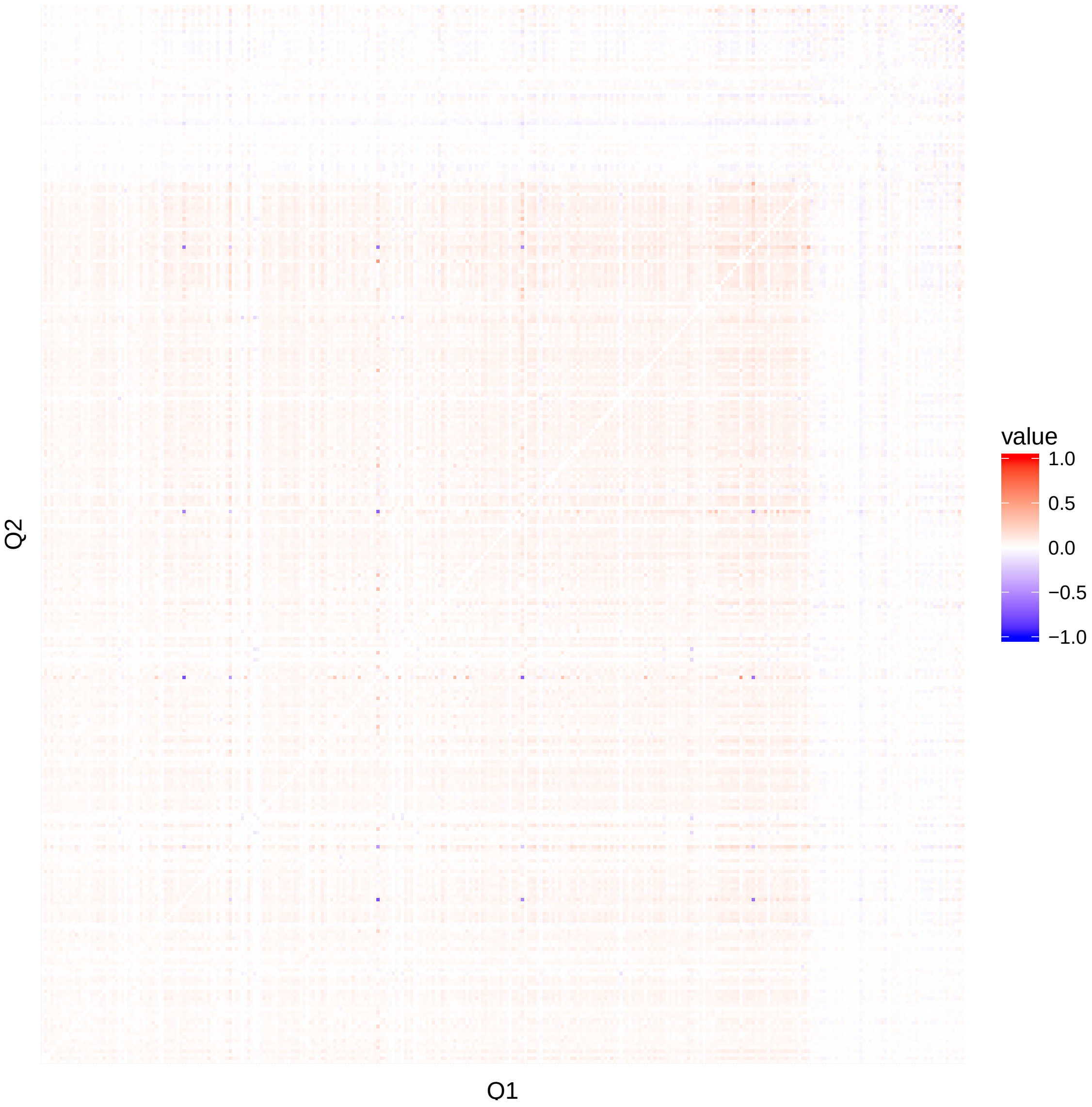

Figure 2.6: Difference between Partial Rank Correlations for Patients with and without Diabetes

Chapter 3

Conclusions

3.1 Summary

In this thesis, we applied a new method to assess covariate-adjusted rank correlations among all pairwise combinations of the most frequent 250 ICD codes and 50 lab results among 472,570 patients in the VUMC EHR. With the combination of cumulative probability models and probability scale residuals, our estimators are rank-based, robust, and can be implemented to large datasets. We highlight some computational challenges in practice and workarounds for applications in big data.

We present our results with estimates and confidence intervals for the partial rank correlation as well as the conditional partial correlation stratified by sex and diabetes status. With further investigations, this could serve as a guidebook for researchers interested in internal associations among EHR features.

The main strength of our approach is that it can be applied to most data types in the EHR while adjusting for covariates and does not require selecting proper transformation for outcomes with skewed distributions. For investigators interested in the difference in conditional correlation across strata, our method can provide a formal hypothesis test. There are a few limitations in our approach as well. First, different lab test results may have various kinds of skewed distributions. We simply used the logit link for all labs in the CPM which has been seen in prior work to be fairly robust [10]. However, one could also use other link functions. Second, we tried our best not to drop useful information in the process of converting lab values to continuous or ordered categorical variables. But information still may be lost for some labs, especially those with text records. Third, the simplified approach in decoding ICD code status (never vs. any) and lab values (taking the earliest one if multiple) may be over simplified. In fact, the longitudinal history for a patient's ICD code assignment and ordered labs may provide some useful information, which calls for more thorough consideration in this procedure.

3.2 Future Research

There are a few potential directions for future work.

First, the data we received contain both ICD-9 and ICD-10 codes. We treated each unique ICD code separately and did not pre-cluster codes. It may be worthwhile to group similar ICD codes before further analyses. For example, there are 5 ICD codes related to diabetes out of the 250 ICD codes we analyzed. We may want to group them into a single ICD code.

Second, the workarounds we proposed in this work are somewhat ad-hoc and warrant further investigation. We did show that the impact was minor when we group lab values more compactly (from 200 bins to 100 bins). It may be worthwhile to study characteristics of those correlations whose inference is greatly affected by binning and to explore how to improve inference for those correlations.

Third, there were reasonable amount of censored observations in lab results. Our quick treatment was rounding them to the nearest detection limit. We actually did a good job when the detection limit was fixed. However, when we have multiple detection limits (e.g. they evolve with time), there are potentially better solutions for censored data. One may want to redo the analyses for those lab values with multiple detection limits to account for censoring by applying different models (e.g. Cox models).

Appendix A

Summary of ICD Codes

Table A.1: Summary of ICD codes

ICD code	percentage	ICD code	percentage	ICD code	percentage	ICD code	percentage	ICD code	percentage
726.10	2.5	781.2	2.9	702.0	3.5	327.23	5.1	785.1	7.7
V58.65	2.5	412	2.9	28	3.5	733.90	5.1	782.3	7.9
719.49	2.6	278.01	2.9	715.90	3.5	Z01.419	5.1	278.00	7.9
787.03	2.6	380.4	2.9	402	3.5	427.9	5.1	461.9	7.9
M54.5	2.6	276.2	2.9	553.3	3.6	V58.61	5.1	724.5	8
728.85	2.6	616.10	2.9	V72.81	3.6	366.16	5.2	V76.51	8
793.19	2.6	477.0	2.9	959.01	3.6	250.02	5.2	780.4	8.1
V10.83	2.6	V45.81	2.9	625.9	3.7	V74.5	5.2	V76.2	8.1
346.90	2.6	478.0	3	721.3	3.7	780.6	5.2	719.41	8.2
789.07	2.6	784.2	3	E11.9	3.7	458.9	5.2	723.1	8.5
R05	2.6	286.9	3	733.00	3.7	584.9	5.2	787.91	8.5
789.59	2.6	V58.67	3	518.81	3.7	211.3	5.4	V72.9	8.7
786.07	2.6	V58.11	3	V05.9	3.8	Z00.00	5.4	300.00	9.2
719.40	2.6	V58.49	3	783.21	3.8	428.0	5.5	244.9	9.2
V12.72	2.6	599.70	3	V22.1	3.8	285.1	5.6	719.46	9.4
470	2.6	V24.2	3	780.39	3.8	466.0	5.7	305.1	9.6
719.43	2.7	E03.9	3	I25.10	3.8	99.04	5.7	465.9	9.7
599.7	2.7	585.9	3	V27.0	3.9	790.29	5.8	272.0	9.8
433.10	2.7	716.90	3.1	789.06	3.9	780.2	5.8	477.9	10.4
558.9	2.7	V22.0	3.1	338.18	3.9	Z79.899	5.8	272.2	10.5
799.02	2.7	V06.8	3.1	780.09	3.9	238.2	5.9	414.01	10.6
V72.3	2.7	724.4	3.1	796.2	3.9	268.9	5.9	285.9	10.7
455.0	2.7	411.1	3.1	V06.1	3.9	V58.81	6	311	10.8
786.52	2.7	780.50	3.1	719.44	3.9	486	6.1	V67.00	10.9
25	2.7	413.9	3.1	V76.44	3.9	V03.82	6.1	518.0	11.2
V77.1	2.7	V54.89	3.2	787.20	4	787.02	6.1	786.09	11.2
263.9	2.7	381.81	3.2	702.19	4.1	V15.82	6.1	786.59	11.9
V58.78	2.7	780.57	3.2	780.97	4.1	788.1	6.1	I10	12.1
789.04	2.7	356.9	3.2	375.15	4.2	562.10	6.1	784.0	12.1
388.70	2.7	578.1	3.2	959.8	4.2	564.00	6.2	V71.7	12.1
626.2	2.7	424.1	3.2	593.9	4.2	518.89	6.3	599.0	12.2
564.1	2.7	473.9	3.2	288.60	4.3	511.9	6.3	250.00	12.8
429.9	2.7	280.9	3.2	627.2	4.3	782.0	6.4	724.2	13.2
472.0	2.7	789.01	3.2	785.6	4.3	429.3	6.4	77	14
724.02	2.8	793.80	3.2	E78.5	4.3	729.1	6.4	789.00	14.1
V70.7	2.8	715.96	3.2	V58.82	4.4	780.60	6.5	401.1	14.5
571.8	2.8	626.4	3.3	496	4.5	719.45	6.5	V72.31	14.8
272	2.8	287.5	3.3	427.89	4.5	427.31	6.5	786.2	15.1
783.1	2.8	V58.66	3.3	E78.2	4.6	V77.91	6.7	V67.09	15.6
V72.62	2.8	715.16	3.3	38.93	4.6	414.00	6.8	V04.81	15.6
88.56	2.8	794.8	3.3	722.52	4.6	719.47	6.9	530.81	15.7
728.87	2.8	276.7	3.4	424.0	4.6	462	6.9	729.5	16.3
414.9	2.8	789.03	3.4	493.90	4.7	276.8	7.1	786.05	16.3
592.0	2.8	Z12.31	3.4	729.81	4.7	782.1	7.2	V58.69	18.8
607.84	2.8	367.1	3.4	790.6	4.8	789.09	7.3	272.4	19
338.29	2.8	K21.9	3.4	780.52	4.8	V45.89	7.4	786.50	19.4
250.01	2.8	276.1	3.4	959.7	4.8	Z23	7.4	V70.0	19.6
279.3	2.8	478.19	3.4	788.41	4.9	785.0	7.5	V72.83	20.4
794.31	2.9	V71.89	3.4	692.9	4.9	V76.12	7.5	780.79	24.3
V45.82	2.9	611.72	3.5	276.51	4.9	787.01	7.5	401.9	31.8

End

Appendix B

Discussion on Conditional Correlation

Assume that we are interested in the partial rank correlations between biomarker Y_1 and Y_2 conditional on diabetes status. The correct approach is to fit one model for Y_1:

$$\log \frac{P(Y_1 \leq k)}{1 - P(Y_1 \leq k)} \sim \mathrm{DIABETES} + \mathrm{SEX} + \mathrm{RACE} + \mathrm{rcs}(\mathrm{BMI}, 3) + \mathrm{rcs}(\mathrm{AGE}, 3) + \beta_{k0},$$

and another model for Y_2:

$$\log \frac{P(Y_2 \leq k^*)}{1 - P(Y_2 \leq k^*)} \sim \mathrm{DIABETES} + \mathrm{SEX} + \mathrm{RACE} + \mathrm{rcs}(\mathrm{BMI}, 3) + \mathrm{rcs}(\mathrm{AGE}, 3) + \beta_{k^*0},$$

and then calculate the sample correlation of the PSRs of these two models. In practice, this correlation is close to the partial correlation that we calculated earlier based on the two models including demographic covariates only (i.e., not including diabetes). For example, the correct conditional correlations between disorders of bursae and tendons in shoulder region (ICD code "726.10") and long-term use of steroids (ICD code "V58.65") are 0.0215 and 0.0224 for diabetes and non-diabetes respectively, and the ones we obtained from models not including diabetes are 0.0243 and 0.0226 respectively. Therefore, instead of refitting new models including diabetes as an extra covariate, we inherited all the outputs from models not including diabetes and used them to obtain the point estimate and inference of conditional correlation in both stratum (diabetes and non-diabetes).

REFERENCES

[1] https://www.healthit.gov/providers-professionals/faqs/what-electronic-health-record-ehr.

[2] Rajiur Rahman and Chandan K Reddy. Electronic health records: A survey. *Healthcare Data Analytics*, 36:21, 2015.

[3] Sima Ajami and Razieh Arab-Chadegani. Barriers to implement electronic health records (ehrs). *Materia socio-medica*, 25(3):213, 2013.

[4] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[5] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[6] Shawn Murphy, Susanne Churchill, Lynn Bry, Henry Chueh, Scott Weiss, Ross Lazarus, Qing Zeng, Anil Dubey, Vivian Gainer, Michael Mendis, et al. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome research*, 19(9):1675–1681, 2009.

[7] Francisco S Roque, Peter B Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Søeby, Søren Bredkjær, Anders Juul, Thomas Werge, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol*, 7(8):e1002141, 2011.

[8] Genevieve B Melton, Simon Parsons, Frances P Morrison, Adam S Rothschild, Marianthi Markatou, and George Hripcsak. Inter-patient distance metrics using snomed ct defining relationships. *Journal of biomedical informatics*, 39(6):697–705, 2006.

[9] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.

[10] Q Liu, BE Shepherd, V Wanga, and C Li. Covariate-adjusted spearman's rank correlation with probability-scale residuals. *Submitted for publication*, 2016.

[11] Qi Liu. *Rank-Based Semiparametric Methods: Covariate-Adjusted Spearman's Correlation with Probability-Scale Residuals and Cumulative Probability Models*. PhD thesis, Vanderbilt University, 2016.

[12] Q Liu, BE Shepherd, C Li, and FE Harrell. Modeling continuous outcomes using ordinal regression with cumulative probabilities, 2017.

[13] https://cran.r-project.org/web/packages/rms/rms.pdf.

[14] Donald A Pierce and Daniel W Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986, 1986.

[15] David R Cox and E Joyce Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):248–275, 1968.

[16] Chun Li and Bryan E Shepherd. A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480, 2012.

[17] Chun Li and Bryan E Shepherd. Test of association between two ordinal variables while adjusting for covariates. *Journal of the American Statistical Association*, 105(490):612–620, 2010.

[18] Bryan E Shepherd, Chun Li, and Qi Liu. Probability-scale residuals for continuous, discrete, and censored data. *Canadian Journal of Statistics*, 44(4):463–479, 2016.

[19] Maurice G Kendall. Partial rank correlation. *Biometrika*, 32(3/4):277–283, 1942.

[20] G Gripenberg. Confidence intervals for partial rank correlations. *Journal of the American Statistical Association*, 87(418):546–551, 1992.

[21] Jeremy Miles and Philip Banyard. *Understanding and using statistics in psychology: A practical introduction*. Sage, 2007.

[22] Leonard A Stefanski and Dennis D Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

[23] S Trent Rosenbloom, William W Stead, Joshua C Denny, Dario Giuse, Nancy M Lorenzi, Steven H Brown, Kevin B Johnson, et al. Generating clinical notes for electronic health record systems. *Appl Clin Inform*, 1(3):232–243, 2010.

[24] http://www.who.int/classifications/icd/en/.