STATISTICAL METHODS FOR PHENOTYPING WITH POSITIVE-ONLY ELECTRONIC HEALTH RECORD DATA

Lingjiao Zhang

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

Jinbo Chen

Jinbo Chen, Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Hongzhe Lee, Professor of Biostatistics

J. Richard Landis, Professor of Biostatistics

Rebecca A. Hubbard, Associate Professor of Biostatistics

Daniel S. Herman, Assistant Professor of Pathology and Laboratory Medicine

STATISTICAL METHODS FOR PHENOTYPING WITH POSITIVE-ONLY ELECTRONIC HEALTH RECORD DATA

ACKNOWLEDGEMENT

I would like to express my deepest gratitude and appreciation to my dissertation advisor, Dr. Jinbo Chen, for her guidance and support thoughout my PhD study. The completion of this dissertation would not be possible without her inspiration and advice. I also thank her for always being caring and encouraging. I would also like to thank my committee members, Dr. Hongzhe Lee, Dr. J. Richard Landis, Dr. Rebecca A. Hubbard, Dr. Daniel S. Herman for their insightful feedback and suggestions for my research.

My sincere thanks goes to the faculty, staff and students in the Department of Biostatistics, Epidemiology & Informatics. I would like to thank my lab rotation advisor, Dr. Sharon Xie, for the invaluable research opportunity. I would also like to thank my fellow students, we have been studying and laughing together, and being there for each other. I would not have enjoyed my journey as much without them. They will always be part of my best memories.

Last and foremost, I would like to thank my family for their unconditional love and support throughout my life. They have always believed in me and encouraged me no matter what, they gave me the strength to reach for the stars and chase my dreams. This thesis is dedicated to them.

ABSTRACT

STATISTICAL METHODS FOR PHENOTYPING WITH POSITIVE-ONLY ELECTRONIC HEALTH RECORD DATA

Lingjiao Zhang

Jinbo Chen

Electronic Health Records-based phenotyping requires fully labeled cases and controls for model training and testing. Due to asymmetric clinical workflow, labeled cases can be much more easily identified than labeled controls. Therefore, data from a group of labeled cases and a large number of unlabeled patients, referred to as "positive-only" data, is frequently accessible with minimum requirement for labeling efforts. This dissertation focuses on statistical methods for training and validating phenotyping models using such positive-only EHR data when the labeled cases can be seen as a representative subset of all cases. In project I, we developed an anchor-variable framework and proposed an accompanying maximum likelihood approach to training a logistic phenotyping model. In project II, we developed a Chi-squared test to assess model calibration through comparing the model-free and model-based estimated number of cases among the unlabeled. We also proposed consistent estimators for predictive performance measures and studied their large sample properties. These methods provide the methodological foundation for positive-only data to be routinely used for training and validating phenotyping models. In project III, we extended the MLE method in project I to accommodate high dimensional predictors by enabling automated feature selection through a proxy phenotype that is available for all patients. We performed extensive simulation studies to assess the performance of the proposed methods and applied them to Penn Medicine EHR data to phenotype primary aldosteronism.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1. Background

The increasing national adoption of Electronic Health Records (EHRs) has the potential to enable implementation of comprehensive, computational clinical decision support and clinical research Hing and Hsiao, 2014; Hripcsak and Albers, 2012; Hsiao and Hing, 2012; Jamoom and Yang, 2016. However, EHRs have been designed primarily to support documentation for medical billing rather than being intricately embedded in clinical diagnostic processes Krenn and Schlossman, 2017; Krousel-Wood et al., 2017; Romano and Stafford, 2011, so that patients' complex, clinical phenotypes are not natively represented in an accurate, precise format Christiansen et al., 2017; Goldstein et al., 2017; Weiskopf and Weng, 2013; Yadav et al., 2018. To overcome this limitation, a variety of EHR phenotyping algorithms Elkan and Noto, 2008; Hong, Liao, and Cai, 2019; Pathak, Kho, and Denny, 2013; Shivade et al., 2013; Yu et al., 2016 have been developed, which are heuristic rules, machine learning techniques and statistical methods based on features extracted from patients' EHRs, both structured and unstructured, to identify patients with a specific phenotype.

1.1.1. Positive-only EHR data

The vast majority of these existing methods require a large, curated dataset of patients who are completely and accurately labeled with regard to the presence or absence of a phenotype. Assigning labels require domain experts to retrospectively review EHR charts and/or prospectively evaluate patients, which takes significant labor and expert knowledge. For many phenotypes, the labor and cost of these processes limit the achievable sample size, which typically ranges from 400 to 600 (Ananthakrishnan et al., 2013; Bejan et al., 2012; Carroll, Eyler, and Denny, 2011; Carroll et al., 2012; Kumar et al., 2014; Liao et al., 2015, 2010; Love, Cai, and Karlson, 2011; Xia et al., 2013; Yu et al., 2017), compromising the accuracy of potential phenotyping models. Moreover, clinical practice workflows are often not symmetric with regards to cases and controls, with only specific patients actively evaluated based on clinical suspicion. There are very few phenotypes for which everyone is actively screened, so clinical practice data is frequently insufficient to identify a large set of gold-standard controls. As a result, for many phenotypes, the most readily accessible

annotations are an incomplete set of gold-standard cases and few or no gold-standard controls. It is highly desirable to develop new phenotyping methods that efficiently and accurately leverage such incomplete phenotyping, or "positive-only" data.

One application for which a phenotyping method that learns from positive-only design is advantageous is identifying patients with and estimating the prevalence of primary aldosteronism (PA). PA is the most common cause of secondary hypertension, thought to affect $\sim 5\%$ of hypertensive patients and up to 20% in specific subgroups of hypertension patients, including hypertension with hypokalemia and resistant hypertension (Oenolle et al., 1993; Rossi et al., 1996; Shigematsu et al., 1997). PA can be treated effectively by unilateral adrenalectomy or targeted medications. Previous methodological work has tried to improve the diagnostic evaluation for PA (Ducher et al., 2015; Lazzarini et al., 2015; Rossi et al., 1998). Unfortunately, PA is not recognized or not treated optimally in many affected patients (Mulatero et al., 2016), so methods for EHR phenotyping have the potential to dramatically improve care for these patients. To develop phenotyping models for PA, a set of cases are readily available because of an existing expert-curated PA research registry (Wachtel et al., 2016) in which every patient was confirmed as a definitive PA case, while it would cost significant efforts to identify a large number of controls. Our goal is to develop new methods for phenotyping model development and validation by efficiently leveraging such "positive-only" data, which is a cohort sample including a group of labeled cases and a large number of unlabeled patients.

1.1.2. Anchor variable framework

The key assumption to our method development is that the labeled cases in "positive-only" data are a random subset of all cases, which can be specified using an "anchor variable" (Halpern et al., 2014). An anchor variable is a binary variable summarizing clinical domain expertise for classifying patients' phenotype. By definition, an anchor variable has perfect positive predictive value (PPV), but is not required to have high sensitivity. That is, anchor positivity indicates presence of the phenotype, but anchor negativity is non-deterministic of the true phenotype status. The second requirement for an anchor variable is that its sensitivity is independent of all phenotype model predictors. An ideal anchor variable is a structured data element in the EHR that is only present in cases, such as the result of a diagnostic confirmatory test or an order that only follows a definitive diagnosis. For example, a pathologic diagnosis of cancer will in most scenarios have a very high PPV, but perhaps an imperfect sensitivity because of variability in practice or documentation,

variability in diagnostic categories, or data incompleteness. For many phenotypes, such definitive diagnostic information may not be available, so surrogates such as diagnosis codes, medications, or note concepts must be considered. Expert knowledge is necessary to select a variable or a composite variable that meets the high PPV and predictor independence requirements (Halpern et al., 2014).

To our best knowledge, the only existing algorithm for learning with positive anchors was proposed by Elkan and Noto (Elkan and Noto, 2008). Their method predicted the probability of phenotype presence through estimating the probability that a subject is anchor positive, motivated by a lemma that the two probabilities differ by a constant factor, anchor sensitivity. This method trains a classifier for the anchor variable in a random subset of patients, then applies the classifier on the rest of patients to estimate the anchor sensitivity. Unfortunately, it yields a consistent estimate of phenotype prevalence only when the predictor distributions for cases and controls are completely separable (Ramaswamy, Scott, and Tewari, 2016), and the estimated phenotype probability may fall outside of the [0,1] range otherwise.

1.2. Method development

In Chapter 2, using a logistic regression model to relate the binary outcome and risk predictors, we propose a maximum likelihood approach to developing a phenotyping model using anchor-labeled cases and unlabeled patients. We compare performance for phenotyping using positive-only data between our method, algorithm proposed by Elkan & Noto, and naive logit, which simply fits a logistic regression model to the positive-only data, treating the unlabeled patients as controls, through extensive simulation studies and development of phenotyping models for hypokalemia requiring oral supplementation amongst primary care patients with hypertension using a simulated anchor variable. We also apply the proposed method to develop phenotyping models for PA using real-world EHRs.

In Chapter 3, we propose novel statistical methods for validating risk prediction models using positive-only EHR data within the anchor variable framework. To overcome the challenge caused by the absence of gold-standard labels for the unlabeled patients, we first propose a model-free estimator for the number of cases in each risk subgroup among the unlabeled patients, taking advantage of the two defining characteristics of the anchor variable. The model-free and model-based (i.e. “expected”) estimated numbers of cases are expected to be close when the fitted model

is correct. Therefore the extent to which these two estimates disagree reflects the severity of the misspecification of the fitted model. We then construct a novel calibration statistic quantifying the disagreement across risk subgroups of the unlabeled patients. We derive the asymptotic distribution of the proposed statistic and show that it follows a χ^2 distribution when the fitted model is correct. We also demonstrate how to estimate the calibration slope using positive-only data. We provide consistent estimators for predictive accuracy measures including true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), negative predictive value (NPV) and area under the ROC curve (AUC), and show their asymptotic properties. We conduct extensive simulation studies to assess the type I error rate of the proposed calibration statistic, to demonstrate its power for detecting model misspecification, as well as to assess statistical consistency of the proposed estimators for predictive accuracy measures. Lastly we apply the proposed validation methods to assess calibration and predictive performances of the preliminary PA phenotyping models that were developed in Chapter 2.

In Chapter 4, we extend the likelihood approach developed in Chapter 2 to a two-stage strategy to incorporate automated feature selection. Although the method proposed in Chapter 2 spares the extensive human resource needed for annotating the training set, it is based on informative features manually selected by clinical domain experts, thus is suitable for developing phenotyping models when the number of potential predictors is relatively small, and not scalable to massive EHR cohorts. Feature selection is one of the main steps and a primary key to success in the development of phenotyping models. Manual variable selection is time consuming and not ideal for high dimensional data. Therefore, various techniques have been applied to reduce the dimensionality space, such as LASSO (least absolute shrinkage and selection operator) or adaptive LASSO (Tibshirani, 1996; Zou, 2006). Yet, these techniques rely on the availability of the gold-standard labeled training set, and its accuracy is oftentimes compromised by the limited sample size.

For many diseases, the screening decision (Yes or No) is informative of the corresponding disease risk, since in practice, the screening tests are only prescribed to individuals who are suspected to be at risk for a particular disease, making it a potential risk factor for disease presence. Although whether a patient has received a screening test or not is oftentimes readily available in the EHR data, instead of directly using the indicator for screening test as a predictor for disease risk, we propose to first build a risk prediction model for the screening decision and then include the pre-

dicted probability of being screened into the risk prediction model for the interest disease. This is because for each patient the screening decision is made by his/her physician based on this particular physician's knowledge and evaluation of the patient's physical condition, including clinical presentation, medical history, family history, laboratory results etc.. Thus the screening decisions varies among physicians and across institutions. Fitting a risk prediction model for the screening decision using the tremendous amount of information contained in EHRs will potentially justify the variance and better identify the ones that should be screened by learning from so many physicians' decision making to patients with similar conditions.

Hereby we propose a two-stage automated feature selection method for developing a risk prediction model for interest disease using positive-only EHR data. In the first stage, we apply penalized logistic regression, such as LASSO or adaptive LASSO, to build a risk prediction model for the easily available screening decision, which performs automated variable selection among the large number of EHR data. In the second stage, we include the predicted probability of screening as a risk factor, along with a short list of expert-hand-picked risk factors, and apply the previously proposed maximum likelihood based anchor learning method to develop the disease risk model. The resultant reduction in the need for manual variable selection would yield improved phenotyping efficiency and scalability. Another benefit of such two-stage phenotyping method is that we can build a robust risk prediction model for screening test in the first stage given the large number of EHR data, which can be applied to predict patients that should be screened, while simultaneously identifying previously overlooked at-risk patients. It could be particularly beneficial if the screening test is expensive or invasive, for example, genetic testing. We conduct extensive simulation studies to demonstrate the performance of our proposed method based on positive-only data, and compare to that of the conventional LASSO and adaptive LASSO performed on an annotated training set. Lastly we apply the proposed two-stage method to build a risk prediction models for primary aldosteronism using data derived from Penn Medicine EHRs.

CHAPTER 2

A MAXIMUM LIKELIHOOD APPROACH TO ELECTRONIC HEALTH RECORD PHENOTYPING USING POSITIVE AND UNLABELED PATIENTS

2.1. Abstract

Phenotyping patients using electronic health record (EHR) data conventionally requires labeled cases and controls. Assigning labels requires manual medical chart review and therefore is labor intensive. For some phenotypes, identifying gold-standard controls is prohibitive. We developed an accurate EHR phenotyping approach that does not require labeled controls. Our framework relies on a random subset of cases, which can be specified using an anchor variable that has excellent positive predictive value and sensitivity independent of predictors. We proposed a maximum likelihood approach that efficiently leverages data from the specified cases and unlabeled patients to develop logistic regression phenotyping models, and compare model performance to existing algorithms. Our method outperformed the existing algorithms on predictive accuracy in Monte Carlo simulation studies, application to identify hypertension patients with hypokalemia requiring oral supplementation using a simulated anchor, and application to identify primary aldosteronism patients using real-world cases and anchor variables. Our method additionally generated consistent estimates of two important parameters, phenotype prevalence and the proportion of true cases that are labeled. Upon identification of an anchor variable that is scalable and transferable to different practices, our approach should facilitate development of scalable, transferable, and practice-specific phenotyping models. Our proposed approach enables accurate semi-automated EHR phenotyping with minimal manual labeling and therefore should greatly facilitate EHR clinical decision support and research.

2.2. Introduction

The adoption of Electronic Health Records (EHRs) by healthcare systems has the potential to enable implementation of comprehensive, computational clinical decision support and clinical research (Hing and Hsiao, 2014; Hripcsak and Albers, 2012; Hsiao and Hing, 2012; Jamoom and Yang, 2016). However, EHRs have been designed primarily to support documentation for medical billing rather than being intricately embedded in clinical diagnostic processes (Krenn and Schlossman, 2017; Krousel-Wood et al., 2017; Romano and Stafford, 2011), so that patients' complex,

clinical phenotypes are not natively represented in an accurate, precise format (Christiansen et al., 2017; Goldstein et al., 2017; Weiskopf and Weng, 2013; Yadav et al., 2018). To overcome this limitation, a variety of heuristic rules and statistical methods have been developed for phenotyping patients using EHR data (Hong, Liao, and Cai, 2019; Pathak, Kho, and Denny, 2013; Shivade et al., 2013; Yu et al., 2016). The vast majority of these existing methods require a large, curated dataset of patients who are completely and accurately labeled with regard to the presence or absence of a phenotype. Such methods require experts to retrospectively review EHR charts and/or prospectively evaluate patients. For many phenotypes, the labor and cost of these processes limit the achievable sample size, compromising the accuracy of potential phenotyping models.

Such approaches can be improved upon by appreciating that clinical practice workflows are often not symmetric with regards to cases and controls. In practice, for most diseases, only specific patients are actively evaluated based on clinical suspicion. There are very few phenotypes for which everyone is actively screened, so clinical practice data is frequently insufficient to identify a large set of gold-standard controls. As a result, for many phenotypes, the most readily accessible annotations are an incomplete set of gold-standard cases and few or no gold-standard controls. It is highly desirable to develop new phenotyping methods that efficiently and accurately leverage such incomplete phenotyping, or "positive-only" data.

One application for which a phenotyping method that learns from positive-only design is advantageous is identifying patients with and estimating the prevalence of primary aldosteronism (PA). PA is the most common cause of secondary hypertension, thought to affect $\sim 5\%$ of hypertensive patients and up to 20% in specific subgroups of hypertension patients, including hypertension with hypokalemia and resistant hypertension (Oenolle et al., 1993; Rossi et al., 1996; Shigematsu et al., 1997). PA can be treated effectively by unilateral adrenalectomy or targeted medications. Previous methodological work has tried to improve the diagnostic evaluation for PA (Ducher et al., 2015; Lazzarini et al., 2015; Rossi et al., 1998). Unfortunately, PA is not recognized or not treated optimally in many affected patients Mulatero et al. (2016), so methods for EHR phenotyping have the potential to dramatically improve care for these patients.

Over the past decade, several methods have been proposed in the machine learning and biostatistics literature for analyzing data arising from such "positive-only" design (Agarwal et al., 2016; Banda et al., 2017; Denis, Gilleron, and Letouzey, 2005; Elkan and Noto, 2008; Liu and Tao, 2015;

Ward et al., 2009). Under this design, one special case with extremely desirable consequences is when the labeled cases are a random subset of the full set of cases. This set of cases could be identified through an active labeling process as part of existing clinical care or research. However, a more generalizable strategy that has the potential to decrease the requirement for manual chart review is the use of binary "anchor variables" Halpern et al. (2014) that summarize clinical domain expertise for classifying patients' phenotype. By definition, an anchor variable has perfect positive predictive value (PPV), but is not required to have high sensitivity. That is, anchor positivity indicates presence of the phenotype, but anchor negativity is non-deterministic of the true phenotype status. The second requirement for an anchor variable is that its sensitivity is independent of all phenotype model predictors. An ideal anchor variable is a structured data element in the EHR that is only present in cases, such as the result of a diagnostic confirmatory test or an order that only follows a definitive diagnosis. For example, a pathologic diagnosis of cancer will in most scenarios have a very high PPV, but perhaps an imperfect sensitivity because of variability in practice or documentation, variability in diagnostic categories, or data incompleteness. For many phenotypes, such definitive diagnostic information may not be available, so surrogates such as diagnosis codes, medications, or note concepts must be considered. Expert knowledge is necessary to select a variable or a composite variable that meets the high PPV and predictor independence requirements (Halpern et al., 2014).

The algorithm for learning with positive anchors proposed by Halpen et al.1was initially introduced by Elkan and Noto (Elkan and Noto, 2008). Their method predicted the probability of phenotype presence through estimating the probability that a subject is anchor positive, motivated by a lemma that the two probabilities differ by a constant factor, anchor sensitivity. This method trains a classifier for the anchor variable in a random subset of patients, then applies the classifier on the rest of patients to estimate the anchor sensitivity. Unfortunately, it yields a consistent estimate of phenotype prevalence only when the predictor distributions for cases and controls are completely separable Ramaswamy, Scott, and Tewari, 2016, and the estimated phenotype probability may fall outside of the [0,1] range otherwise.

To enable more efficient and accurate EHR phenotyping and estimation of phenotype prevalence in the setting of incomplete clinical training phenotypes, we hereby propose a maximum likelihood method to develop a logistic regression prediction model using positive-only EHR data using cases

identified by chart review and/or anchor variables. We have demonstrated, via extensive simulation studies, development of phenotyping models for hypokalemia requiring oral supplementation amongst primary care patients with hypertension using a simulated anchor variable, and development of a PA phenotyping model for patients screened for PA using two different real-world sets of cases, that this method develops models that accurately identify unlabeled cases and yields consistent estimate of phenotype prevalence.

2.3. Methods

2.3.1. Positive-only data

Let Y denote the latent binary label for the phenotype (1: case; 0: control), $\mathbf{X}$ denote the predictor variables of Y, and S denote the binary anchor variable (1: positive; 0: negative). Here $(\mathbf{X}, Y, S)$ are considered as random variables from which EHR patients, including both anchor-positive cases and unlabeled patients, are randomly drawn, with only $(\mathbf{X}, S)$ observed. For a well-chosen anchor variable, Y takes value 1 whenever $S = 1$, i.e. $p(Y = 1|S = 1) = 1$. But Y can take either value 1 or 0 when $S = 0$. The anchor sensitivity being independent of $\mathbf{X}$ can be formalized as (Elkan and Noto, 2008)

$$p(S = 1|Y = 1, \mathbf{X}) = p(S = 1|Y = 1) \equiv c, \tag{2.1}$$

where c is a constant between 0 and 1. We use a logistic working model to relate Y and bx which is commonly implemented in the EHR setting, although our method is applicable for any parametric model that is reasonable for modeling binary outcome variables:

$$\text{logit } p(Y = 1|\mathbf{X}; \boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}. \tag{2.2}$$

Here we allow $\mathbf{X}$ to include a vector of 1 so that the intercept parameter is implicitly included in the logit function $\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$. For notational simplicity, we use $P(\mathbf{X}; \boldsymbol{\beta})$ to denote $p(Y = 1|\mathbf{X}; \boldsymbol{\beta})$. Let $F(X)$ denote the cumulative distribution function of $\mathbf{X}$, $f(\mathbf{X})$ the corresponding probability density function, and q the phenotype prevalence, $q = p(Y = 1) = \int P(\mathbf{X}; \boldsymbol{\beta})dF(\mathbf{X})$. Let h be the probability of anchor being positive, $h = p(S = 1)$. It is easy to show that anchor sensitivity $c = h/q$ by applying (2.1).

2.3.2. Algorithms for learning with positive-only data

In this section we describe two methods for fitting model (2.2) and estimating parameters c and q from positive-only data, an algorithm developed by Elkan and Noto (Elkan and Noto, 2008; Halpern et al., 2014), which to our best knowledge is the only available method to date ("EN algorithm"), and a maximum likelihood method that we here propose ("ML method"). We also describe a "Naive" logistic modeling referred to as Naive logit, which simply treats the unlabeled patients as controls and has largely been the standard method for analyzing positive-only data.

EN algorithm

(Elkan and Noto, 2008; Halpern et al., 2014) This algorithm learns from positive-only data by building upon a lemma relating the models for predicting outcome status Y and anchor status S, $p(Y = 1|\mathbf{X}) = p(S = 1|\mathbf{X})/c$. This method requires the positive-only data to be randomly separated into two parts, the training set and validation set. It can be summarized into three steps. First, a logistic regression model is developed for predicting anchor status S, $g(X) = p(S = 1|\mathbf{X})$, by fitting the model directly to the training set with S being the outcome variable. Second, the anchor sensitivity c is estimated using the validation set. They proposed that c be estimated as $\hat{c} \equiv \sum_{i=1}^{n^v} g(\mathbf{X}_i)I(S_i = 1)/\sum_{i=1}^{n^v} I(S_i = 1)$ with n^v being the size of the validation set. Third, the probability of the outcome status is estimated for an unlabeled patient as $p(Y = 1|\mathbf{X}) = p(S = 1|\mathbf{X})/\hat{c}$. The algorithm is intuitive and easy to implement. But the proposed estimator of c is often biased unless the predictor distribution for cases and controls are completely separable (Ramaswamy, Scott, and Tewari, 2016), thereby leading to biased estimation of $p(Y = 1|\mathbf{X})$ and prevalence q. When the estimated c is biased towards 0, the estimated $p(Y = 1|\mathbf{X})$ could exceed 1.

ML method

We hereby propose to use the maximum likelihood ("ML") method for fitting model $p(Y = 1|\mathbf{X}; \boldsymbol{\beta})$ and estimating anchor sensitivity c simultaneously. The likelihood function for the observed data for all N patients is as follows,

$$\begin{aligned} L(\boldsymbol{\beta}, c) &= \prod_{i=1}^{N} p(\mathbf{X}_i, S_i = 1)^{S_i} \times p(\mathbf{X}_i, S_i = 0)^{1-S_i} \\ &\propto \prod_{i=1}^{N} \{cP(\mathbf{X}_i; \boldsymbol{\beta})\}^{S_i} \times \{1 - cP(\mathbf{X}_i; \boldsymbol{\beta})\}^{1-S_i}. \end{aligned}$$

As shown in Appendix I, $(\boldsymbol{\beta}, c)$ are identifiable with positive-only data. Thus we can obtain the ML estimates $(\widehat{\boldsymbol{\beta}}, \hat{c})$ by maximizing the log likelihood function $logL(\boldsymbol{\beta}, c)$. The large sample variance-covariance matrix of these estimates can be established from the inverse of the information matrix. We propose two methods for estimating phenotype prevalence q. Because q can be expressed as h/c, it can be estimated as $\hat{q} = \hat{h}/\widehat{c}$, where $\hat{h}$ is the maximum likelihood estimate of $p(S = 1)$ and equal to the sample fraction of those with $S = 1$. Alternatively, it can be estimated as the average of the estimated phenotype probabilities, $\hat{q} = N^{-1} \sum_{i=1}^{N} P(\mathbf{X}; \widehat{\boldsymbol{\beta}})$. In Appendix II, we described an extension of this method by allowing c to vary with respect to a small number of pre-specified strata.

Naive logit

This method simply fits a logistic regression model to the positive-only data, treating the unlabeled patients as controls. That is, the anchor status S was used as if it were the truth label Y.

2.4. Simulation Studies

We carried out extensive simulation studies to evaluate the performance of the ML method relative to the EN algorithm and naive logit for phenotyping using positive-only data. We also included results from the Ideal learning where fully labeled data is used to fit standard logistic regression models with the true labels Y as the outcome variable. The Ideal learning, feasible only for simulation studies, was used as the comparison benchmark. Given that anchor-positive patients are classified as cases by definition, we assess classification accuracy only among the unlabeled patients, using measures true positive rate (TPR_v), false positive rate (FPR_v), positive predictive value (PPV_v), negative predictive value (NPV_v) at a decision threshold v and area under the ROC curve (AUC). We also demonstrated statistical consistency for the ML estimators $\widehat{\boldsymbol{\beta}}$, $\hat{c}$ and $\hat{q}$ across a range of prevalence and anchor sensitivities in Appendix IV.

2.4.1. Simulation Settings

To mimick the complex data structure in EHR, we generated the binary outcome variable Y from a logistic regression model with 9 predictors,

$$\text{logit } p(Y = 1|\mathbf{X}; \boldsymbol{\beta}) = \beta_0 + \sum_{k=1}^{9} \beta_k X_k, \tag{2.3}$$

with (X_1, X_2, X_3), (X_4, X_5, X_6), (X_7, X_8, X_9) representing weak, moderate, and strong predictors respectively by setting the corresponding parameter coefficients at $(\beta_1, \beta_2, \beta_3) = (0.2, 0.4, 0.6)$, $(\beta_4, \beta_5, \beta_6) = (-1.0, -1.4, 1.8)$, $(\beta_7, \beta_8, \beta_9) = (-2.0, 2.4, 2.8)$. The 9 predictors were independently

distributed, with X_1, X_4 and X_7 generated from normal distribution $N(5, 10)$, X_2, X_5 and X_8 from Bernoulli distribution with success rate 0.5, and X_3, X_6 and X_9 from logit transformed standard uniform distribution. The value of β_0 was varied accordingly to achieve the phenotype prevalence at 5%, 10%, 15% and 20% respectively. The anchor sensitivity c was fixed at 0.5. For each case ($Y = 1$), the anchor variable S was generated according to a Bernoulli distribution with success rate c. For each control ($Y = 0$), S was always set to 0. In each Monte Carlo simulation, we drew a random sample of size 10,000 as the training set, and a disjoint testing set of 5,000. To implement the EN algorithm, 20% of the training set was put aside as the validation set to estimate c. For each parameter combination, we iterated the simulation 1,000 times. Below we focus our discussion on the results when the phenotype prevalence was set as 10%. Results for the other three prevalence values were similar and included in Supplementary Materials.

2.5. Results

2.5.1. Simulation results

As shown in Table 2.1, our ML method yielded consistent estimates of the anchor sensitivity c and the phenotype prevalence q that were nearly identical to that of the Ideal logistic regression using the truth labels. EN algorithm underestimated anchor sensitivity c as 0.37 with empirical standard error (ESE) 0.04, with the bias potentially caused by the overlapping of predictor distributions, leading to overestimation of prevalence q as 0.14 (ESE: 0.01). Among the unlabeled patients, 4% (ESE: 0.5%) of the EN algorithm-predicted probabilities $\hat{p}(Y = 1|\mathbf{X})$ were greater than 1. The EN algorithm estimates remained nearly identical when we increased the size of the validation set to 5000. As demonstrated in Table 2.2 and Figure 2.1, ML method achieved comparable predictive accuracy to that of Ideal learning among the unlabeled patients, with similar AUC (0.994 and 0.994 respectively) and Precision-Recall curve, which is a plot of PPV (y-axis) and TPR (x-axis) for different risk thresholds. EN algorithm and Naive logit achieved identical AUC of 0.993. However, at their respective threshold for 80% sensitivity (TPR), ML method achieved higher specificity, with PPV and FPR estimated as 86% and 0.7%, compared to 84% and 0.9% for EN algorithm and 83% and 0.9% for Naive logit. Similarly, at thresholds for 80% PPV, ML method yielded higher TPR than EN algorithm and Naive logit (Table 2.2, Figure 2.1).

As shown in Table A.1, the averaged ML parameter estimates $(\widehat{\boldsymbol{\beta}}, \hat{c}, \hat{q})$ appeared very close to the true values. The negligibly small biases ($< 4\%$) indicated statistical consistency of the proposed

ML estimators. The variances of the ML estimates can be well approximated by the asymptotic standard errors (ASE) as their average across simulated data and the empirical standard error were very close with difference $< 6\%$. The ML therefore is superior to EN algorithm also because the variance of the estimates $(\widehat{\boldsymbol{\beta}}, \hat{c}, \hat{q})$ can be obtained conveniently.

2.5.2. Method validation using real-world EHR data and a simulated anchor variable

Next we sought to validate the ML method using real-world EHR data and a simulated, mes perfect anchor variable. We compared the performance of the proposed ML method to that of the EN algorithm, Naive logit and Ideal learning for identifying a patient population that should be screened for PA, those with hypokalemia requiring oral supplementation, from amongst 10,000 Penn Medicine primary care patients with hypertension. Patients (≥ 18 years old) were randomly selected from amongst those with ≥ 5 office visits over ≥ 3 distinct years, including ≥ 2 at one specific primary care practice, between 2007 and 2017. The population was restricted to patients with hypertension by filtering for ≥ 2 outpatient encounters with a hypertension diagnosis code (ICD9 401.*, 405.*; ICD10 I10.*, I15.*). We defined the phenotype status Y for each patient as whether the patient had ≥ 3 outpatient orders for oral potassium supplementation. The resulting dataset contained 796 (8%) cases and 9204 controls. As our goal was to demonstrate the performance of the positive-only phenotyping methods when a good anchor variable is available, we artificially created an anchor variable S with sensitivity 0.2 by randomly setting S to 1 for 20% of all cases ($Y = 1$), and to 0 for the remaining 80% of cases and for all controls. The predictors of interest, $\mathbf{X}$, were selected by clinical experts from amongst diagnosis codes, medication prescriptions, laboratory results, vital signs, encounter meta information (Table A.4). We emphasize that ML method and EN algorithm only need S and $\mathbf{X}$ for model training. The true label Y was only used for Ideal learning. For EN algorithm, we randomly set aside 20% of the dataset for estimating anchor sensitivity. We performed 10-fold cross validation for estimating all predictive accuracy metrics based on phenotype status Y. The proposed ML method estimated anchor sensitivity c as 0.17 (ASE: 0.04), which was much closer to the true value 0.2 than the EN estimate of 0.05. Consequently, the ML estimate of prevalence q, 0.09 (ASE: 0.02), was much closer to the true value 0.08 than the EN estimate 0.32.

For phenotyping accuracy, Ideal learning achieved the highest AUC of 0.86, which is the best discrimination since it used the Y truth labels. ML method and Naive logit both achieved AUC as 0.85, which was slightly higher than that of EN algorithm, 0.83. At their respective thresholds for reaching 70% TPR, ML model achieved a PPV of 23%, which was identical to that of Ideal learning

(23%) and Naive logit (23%), and slightly higher than that of EN algorithm (20%) (Figure 2.2). Furthermore, FPR of ML method and Ideal learning were also identical (16%), which were lower than that of EN algorithm (19%). The NPV was comparable for all full methods. At their respective threshold for achieving 50% PPV, ML model achieved a TPR of 44%, which was comparable to that of Ideal learning (45%) and higher than that of Naive logit (38%) and EN algorithm (39%). In addition, 5% of the EN algorithm-predicted probabilities $\hat{p}(Y = 1|\mathbf{X}, S = 0)$ were greater than 1.

2.5.3. A preliminary phenotyping model for primary aldosteronism using real-world predictors and cases

We then applied the proposed ML method to develop a preliminary model for identifying PA patients using a dataset derived from Penn Medicine EHR that contains 6319 patients who had an order for a PA screening laboratory test. Predictor variables, listed in Tables S5 and S6, were selected by clinical experts. Because lab test ordering is non-random and potentially informative of phenotype status, we created binary variables indicative of result presence and included these binary indicators as well as the original lab test variables, with missing values replaced by zero. Missing values in variables other than laboratory results were imputed with respective mean values. In addition, variables with highly positively skewed distributions were log-transformed, and all continuous variables were standardized. Since the focus of this work is on demonstrating the performance of the ML method with a wisely chosen set of cases, we leveraged an existing expert-curated PA research registry(Wachtel et al., 2016) in which every patient was a definitive PA case (Case set A; N=149, 2.4%). The patients were identified for the registry because they underwent a diagnostic procedure, adrenal vein sampling, which in our practice is only performed on patients definitively diagnosed with primary aldosteronism. We later supplemented this set using an anchor variable strategy to include patients with a laboratory test order for adrenal vein cortisol, which is only performed as part of this the adrenal vein sampling procedure ("Set B"). This anchor variable labeled an additional 47 patients, increasing the set of cases to a total of 196 (3.1%). The additional patients included many who were tested after 2015 and some that were missing from the registry for unknown reasons. We evaluated the predictive accuracy of the models using TPR and PPV. TPR was estimated using the case set via 10-fold cross-validation for models developed by each of the three methods. This was plausible because the independence property of anchor variable implied that TPR among all PA cases can be estimated by TPR among anchor-labeled cases, $TPR = p(\hat{p} > v|Y = 1) = p(\hat{p} > v|S = 1)$. To estimate PPV, charts for patients receiving longi-

tudinal care and $\hat{p}(Y = 1|\mathbf{X}) \geq 0.2$ by each model were reviewed by clinician I.A. and ambiguous cases were further reviewed by D.S.H. There were 185 charts reviewed and were adjudicated with respect to PA as 132 positive, 5 unknown, and 48 negative. The final models were included in Table A.7.

Using case set A, the ML method estimated the label sensitivity c as 0.56, compared to 0.35 of EN algorithm. Consequently, it estimated phenotype prevalence q as 4% (95% CI: 3%-5%), compared to 7% by EN. For both case sets, the ML-fitted model appeared to have high discriminatory power, indicated by the double-peaked histogram of $\hat{p}(Y = 1|\mathbf{X})$, with anchor-positive cases mostly having high predicted probabilities (Figure 2.3). ML method achieved consistently higher TPR than the other two methods. For example, at threshold 0.5, TPR was estimated as 0.66, 0.59 and 0.28 for ML method, EN algorithm, and Naive logit respectively (Table 2.3). In addition, 0.6% of the EN algorithm-predicted probabilities $\hat{p}(Y = 1|\mathbf{s}, S = 0)$ were greater than 1. Results using case set B were similar, with c estimated as 0.62 and 0.41 by ML method and EN algorithm, respectively, and q estimated as 5% (95% CI: 4%-6%) by ML method and 8% by EN algorithm. Again, TPR of ML method was consistently higher than that of the other two methods (Table 2.3), and 0.7% of the EN algorithm-predicted probabilities $\hat{p}(Y = 1|\mathbf{X}, S = 0)$ were greater than 1.

According to the chart review results, the models based on case set A and B achieved similar PPVs at threshold 0.2 (75% and 76% respectively) and threshold 0.5 (78% and 77% respectively). Notably, the ML models identified 7 unlabeled PA patients that do not meet PA heuristic rules $\geq$ 2 PA diagnosis codes (ICD9: 255.10, 255.11, 255.12; ICD10: E26.01, E26.02, E26.09, E26.9) or PA laboratory testing results that meet conservative diagnostic criteria (aldosterone $\geq$ 15 ng/dL, plasma renin activity $<$ 0.5 ng/ml/hr, aldosterone:PRA $\geq$ 30).

To assess the sensitivity of our method with respect to the independence requirement of anchor variables, we performed stratified analyses as described in Appendix II by allowing anchor sensitivity to vary with the encounter variables. Specifically, for each variable, we allowed anchor sensitivity to differ between the two strata, above or below median. The phenotype prevalence estimates were nearly identical to those presented earlier where the independence assumption was imposed, and differences between the two stratum-specific anchor sensitivity estimates were minor (Table A.8). The estimated regression coefficients were very similar as well (Tables A.9).

2.6. Discussion

A key step to comprehensive and accurate EHR phenotyping is the development of accurate approaches that efficiently leverage clinical expertise and available data. Currently, to train phenotyping algorithms, most existing methods require a complete set of expert-annotated cases and controls or naively treat unlabeled patients as definitive controls. Our proposed ML method builds accurate model classifiers based upon a random sample of positive cases and a large number of unlabeled patients. To identify a random sample of cases, the method can leverage domain expertise summarized in the form of an anchor variable, with modest upfront effort from clinical experts Halpern et al. (2014). Compared to standard strategies, this method dramatically decreases the need for labor-intensive chart annotation and prospective phenotyping.

Our ML method appeared to consistently outperform EN algorithm and Naive Logit in all numerical studies according to predictive accuracy metrics considered and estimates of anchor sensitivity and phenotype prevalence. Another notable feature of our ML method is transferability to other practices. The anchor concept may itself be more easily transferred rather than the full model Halpern et al. (2016). Model validation with respect to calibration and predictive accuracy classically requires annotated labels for a random set of patients. However the anchor variable framework and ML method allow the development of novel methods for internally assessing model calibration and predictive accuracy using positive-only data (which we are currently working on), precluding the need for external model validation. Thus to generalize our method to secondary sites, chart review need only be performed to confirm that the anchor has very high PPV for the phenotype of interest, which is considerably less burdensome validation of a classically fit or transferred model. Recent work has considered phenotyping methods that take advantage of both noisy labels with random error and anchor variable framework Ducher et al. (2015). Our method can similarly be extended in this regard.

The model performance relies upon the independence of labeled cases, or anchor sensitivity, and model predictors. This necessitates meticulous selection of the anchor based on clinical expertise. If unclear, the appropriateness of an anchor could be supported by explicitly validating the estimated phenotype prevalence, model sensitivity, or the conditional independence assumption. We refer readers to Halpern's work Halpern et al. (2016) for methods to support defining potential anchors. The phenotyping models for PA based on the two case sets were similar, which to a certain extent

suggests the robustness of the ML method with respect to anchor selection. We also extended the current implementation of anchor variables to allow anchor sensitivity to vary across a fixed number of discrete strata that are pre-defined by patient EHR data (Appendix II). We then applied this approach in our PA modeling to exclude major predictor-anchor dependence.

In this work, we applied our method to develop models to identify patients with PA. In selecting as cases PA patients who underwent a subtyping diagnostic procedure, we targeted patients with more severe and actionable disease rather than all PA patients. Thus, based on the conditional independence assumption, our models suggest that amongst patients screened for PA, including primary care and specialty referral populations, the prevalence of PA eligible for adrenal vein sampling is approximately 5%. To validate the prevalence estimation, it would be ideal to have expert annotation for a random sample of patients. Because the prevalence of PA is low, a large number of patients would need to be annotated. Unfortunately, since the diagnosis of PA requires specific diagnosis testing (measurement of blood aldosterone and plasma renin activity), chart review is not sufficient to identify all PA patients in a cohort. Therefore, it is implausible to obtain a sufficient annotated validation set. That said, our prevalence estimates were reasonable based on the literature: $\sim 5\%$ for the prevalence of all PA in primary care populations33, $\sim 10\%$ for prevalence of PA in tertiary care settings Rossi et al. (2006). One recent Dutch study Käyser et al. (2018) that has a similar study design demonstrated a 3% (95% CI: 1.4-4.9) prevalence of PA confirmed by provocative testing among patients newly diagnosed with hypertension and screened for PA.

Our ML method demonstrated good sensitivity and PPV for identifying PA patients. But, as this was merely a proof-of-concept analysis, there is considerable room for improvement. We focused on specific predictors selected by domain experts, and did not exhaustively explore feature selection and engineering. Our current method is suitable for developing phenotyping models when the number of potential predictors is far less than the number of records. It is of interest to explore additional predictors across high dimensional EHR data, which we expect would lead to models with improved accuracy and more precise estimates of anchor sensitivity. We plan to extend our current method to facilitate variable selection in building the prediction model. For example, Yu et al Yu et al. (2016) proposed a surrogate-assisted feature extraction (SAFE) method for high-throughput EHR phenotyping. It reduces the needed number of gold-standard labels by selecting a candidate set of predictors that are correlated with ICD-9 codes and/or natural language processing-extracted

concept counts. Potentially, a SAFE-like approach could be applied to generate candidate predictors for use in our maximum likelihood based anchor learning method. The resultant reduction in the need for manual variable selection would yield improved phenotyping efficiency and scalability. Similarly, we would expect considerable further gains from more extensive modeling, including exploration of alternative missing data approaches.

2.7. Conclusion

The incompleteness and asymmetry of EHR data limits its use for clinical decision support and research. We have developed a novel likelihood-based method that uses labeled cases and unlabeled patients to simultaneously enable accurate model development and identification of unlabeled cases. We expect this method will facilitate phenotype model development and transferability for a wide variety of EHR clinical decision support and research applications.

Table 2.1: Estimate of c and q, presented as mean (empirical standard error) over 1000 iterations.

	Anchor sensitivity c	Phenotype prevalence q
True value	0.5 (-)	0.1 (-)
Ideal learning	-	0.10 (0.0003)
Naive logit	-	0.05 (0.002)
EN algorithm	0.37 (0.04)	0.14 (0.01)
ML method	0.50 (0.021)	0.10 (0.004)

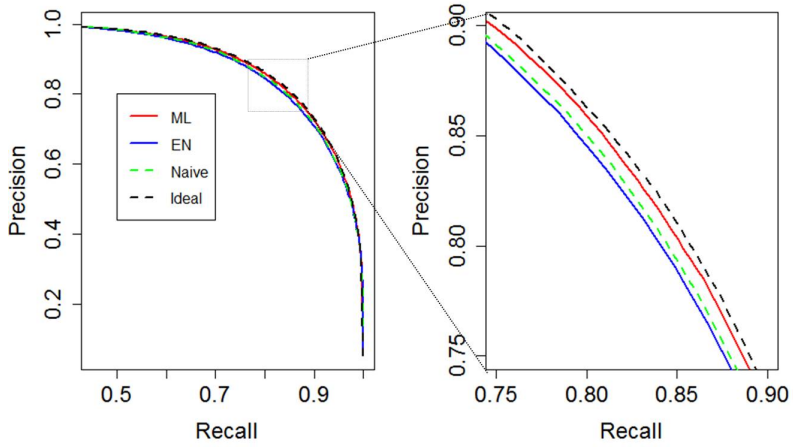

Figure 2.1: Precision-Recall curve of the four methods from simulation study with 10% prevalence.

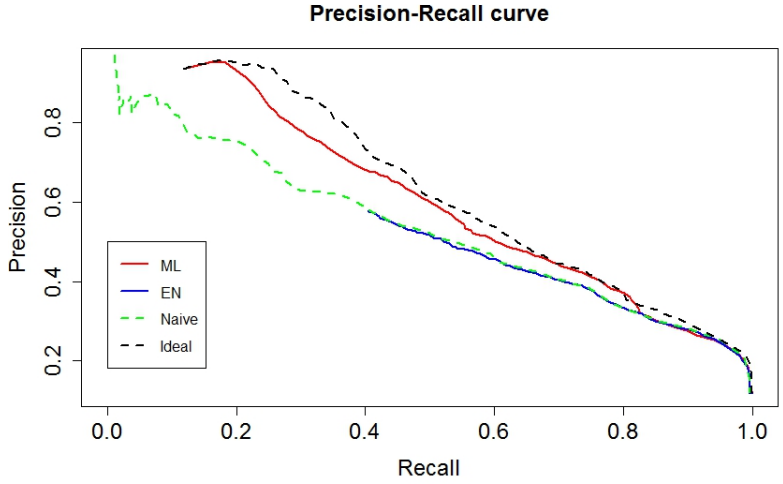

Figure 2.2: Precision-Recall curve of the four methods for identifying patients with hypokalemia requiring potassium supplementation amongst Penn Medicine primary care patients with hypertension.

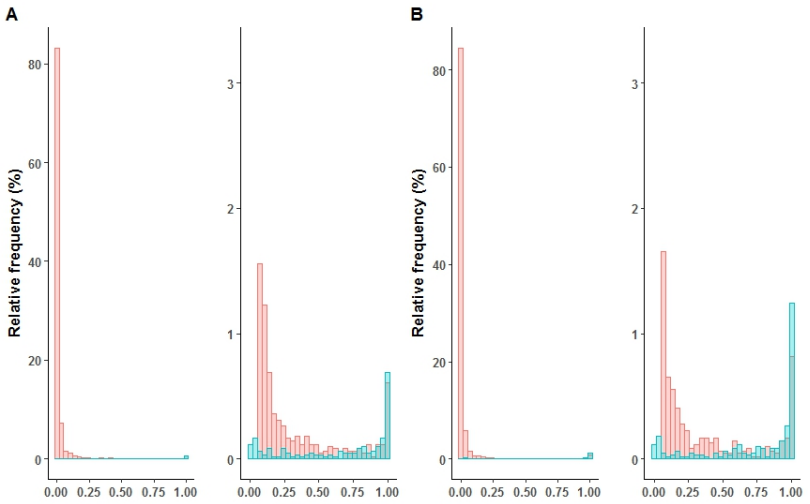

Figure 2.3: Histogram of estimated probabilities of PA, $p(Y = 1|\mathbf{X})$ based on case set A (Panel A) and case set B (Panel B). The left panels display the probability histogram with the range of y-axis being 0-100%. The right panels display the zoomed-in probability histogram with y-axis limited to 0-3%. Labeled cases (blue) and unlabeled patients (red) indicated.

Table 2.2: Phenotyping accuracy measures at thresholds selected to achieve $TPR = 0.8$ (A) or $PPV = 0.8$ (B), presented as mean (empirical standard error) over 1000 iterations.

	(A) TPR=0.8				(A) PPV=0.8			
	Threshold	FPR	PPV	NPV	Threshold	TPR	FPR	NPV
Ideal learning	0.63	0.007 (0.001)	0.86 (0.02)	0.99 (0.002)	0.5	0.90 (0.02)	0.01 (0.002)	0.99 (0.001)
Naive logit	0.15	0.009 (0.002)	0.83 (0.03)	0.99 (0.002)	0.14	0.84 (0.03)	0.01 (0.002)	0.99 (0.002)
EN algorithm	0.42	0.009 (0.003)	0.84 (0.05)	0.99 (0.003)	0.38	0.84 (0.04)	0.01 (0.004)	0.99 (0.002)
ML method	0.63	0.007 (0.002)	0.86 (0.03)	0.99 (0.002)	0.5	0.89 (0.03)	0.01 (0.002)	0.99 (0.002)

Table 2.3: PA model TPR estimated using labeled patients.

	Case Set A			Case Set B		
Threshold	ML method	EN algorithm	Naive logit	ML method	EN algorithm	Naive logit
0.10	0.85	0.85	0.73	0.89	0.86	0.77
0.20	0.79	0.74	0.6	0.85	0.8	0.63
0.30	0.74	0.7	0.49	0.83	0.71	0.56
0.40	0.7	0.64	0.35	0.8	0.64	0.47
0.50	0.66	0.59	0.28	0.78	0.61	0.41
0.60	0.62	0.55	0.26	0.72	0.57	0.34
0.70	0.56	0.48	0.21	0.66	0.55	0.3
0.80	0.52	0.42	0.15	0.6	0.49	0.24

CHAPTER 3

TESTING CALIBRATION OF RISK PREDICTION MODELS USING POSITIVE-ONLY ELECTRONIC HEALTH RECORD DATA

3.1. Abstract

Validation of risk prediction models using Electronic Health Records (EHRs) data conventionally requires gold-standard case and control labels. The labeling process requires clinical experts to retrospectively review patients' medical charts, therefore is labor intensive and time consuming. For some disease conditions, it is prohibitive to identify gold-standard controls because routine clinical assessments are performed for selective patients who are deemed to possibly have the condition. To build a model for phenotyping patients in EHRs, the most readily accessible data is often for a cohort consisting of a set of gold-standard cases and a large number of unlabeled patients. Hereby we propose methods for assessing model calibration and predictive accuracy using such "positive-only" EHR data that do not require gold-standard controls, provided that the labeled cases are representative of all cases. For model calibration, we propose a novel statistic that aggregates differences between model-free and model-based estimated numbers of cases across risk subgroups, which asymptotically follows a Chi-squared distribution. We additionally demonstrate that the calibration slope can also be estimated using such "positive-only" data. We propose consistent estimators for predictive performance measures and derive their large sample properties. We demonstrate performances of the proposed methods through extensive simulation studies and apply them to Penn Medicine EHRs to validate two preliminary models for predicting the risk of primary aldosteronism (PA).

3.2. Introduction

Many efforts have been spent on the development of automated phenotyping algorithms using electronic health record (EHR) data, which are classification rules based on features extracted from patients' EHRs, both structured and unstructured, to infer whether a patient has a specific phenotype (Elkan and Noto, 2008; Hong, Liao, and Cai, 2019; Pathak, Kho, and Denny, 2013; Shivade et al., 2013; Yu et al., 2016). Conventionally, the development and validation of EHR-based phenotyping algorithms rely on a curated dataset, in which every patient is labeled accurately with regard to presence (case) or absence (control) of the phenotype of interest. For example,

conventional Hosmer-Lemeshow type of calibration tests are based on the differences between observed and expected numbers of phenotypes across patient subgroups (Hosmer, Lemeshow, and Klar, 1988; Hosmer Jr, Lemeshow, and Sturdivant, 2013; Song et al., 2014; Tsiatis, 1980; Windmeijer, 1990). Accurate assignment of case and control labels usually requires domain experts to retrospectively review patients' medical charts, therefore is labor intensive and time consuming. This limits the achievable sample size and compromises phenotyping accuracy. In addition, for many diseases, since patients are usually actively screened based on clinical suspicion, the EHR data is often insufficient for identifying a large group of gold-standard controls. Therefore, frequently, most readily accessible EHR data is for an incomplete set of gold-standard cases and a large number of unlabeled patients whose case and control statuses are unknown. Hereafter, we refer to the data for such an incompletely labeled EHR cohort as "positive-only" data

We previously proposed a maximum likelihood approach to developing phenotyping models using positive-only data within an anchor variable framework (Zhang et al., 2019). This approach greatly reduces dependency on gold-standard labeling since it does not require labeled controls. It relies on the specification of an anchor variable (Halpern et al., 2014) for the target phenotype to label the set of cases instead of labor-intensive manual chart reviewing. An anchor variable is a binary variable specified by domain experts as highly informative of patients' true phenotype status. The patients with positive anchor variable values form a random subset of all cases in the population because of two unique properties of the anchor varible. First, anchor variable has perfect positive predictive value, i.e., anchor being positive indicates the presence of the phenotype, while it being negative is nondeterministic of the patient's phenotype status. Second, anchor variable has constant sensitivity, i.e., conditional on phenotype presence, the anchor being positive or negative is independent of phenotype model predictors. Anchor variable specification typically requires less upfront effort from clinical experts compared to labor-intensive chart review for identifying gold-standard cases and controls.

In this project, motivated by avoiding labor intensive labeling for model validation, we propose statistical methods for validating risk prediction models using the positive-only EHR data within the anchor variable framework. To overcome the challenge caused by the absence of gold-standard labels for the unlabeled patients, we first propose a model-free estimator for the number of cases in each risk subgroup among the unlabeled patients, taking advantage of the two defining character-

istics of the anchor variable. The model-free and model-based (i.e. "expected") estimated numbers of cases are expected to be close when the fitted model is correct. Therefore the extent to which these two estimates disagree reflects the severity of the misspecification of the fitted model. We then construct a novel calibration statistic quantifying the disagreement across risk subgroups of the unlabeled patients. We also demonstrate how to estimate the calibration slope using positive-only data with our proposed maximum likelihood approach (Zhang et al., 2019).

In organizing the paper, we first present our statistic for assessing model calibration using positive-only data when phenotype prevalence is known or unknown respectively. We also derive the asymptotic distribution of the proposed statistic and show that it follows a χ^2 distribution when the fitted model is correct. We then show how to estimate the calibration slope using positive-only data. We next present estimates of predictive accuracy measures including true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), negative predictive value (NPV) and area under the ROC curve (AUC), and show their asymptotic properties. We conduct extensive simulation studies to assess the type I error rate of the proposed calibration statistic, demonstrate its power for detecting model misspecification, as well as assess statistical consistency of the proposed estimators for predictive accuracy measures. Lastly we apply the proposed validation methods to assess calibration and predictive performances of the preliminary models for predicting primary aldosteronism that were derived from Penn Medicine EHR data (Zhang et al., 2019).

3.3. Methods

3.3.1. The Anchor variable framework

Let Y_i denote the binary label for the phenotype of interest for patient i, e.g., an indicator for the patient being a case ($Y_i = 1$) or control ($Y_i = 0$). Let $\mathbf{X}_i$ denote a p-dimensional vector of risk predictors for Y_i, which includes 1 to allow the intercept term, and S_i denote the binary anchor variable, which takes values $S_i = 1$ (positive) or $S_i = 0$ (negative). The positive-only data consists of observations on $\mathbf{X}$ and S for a random sample of N patients, $\{(\mathbf{X}_i, S_i), i = 1, \ldots, N\}$, where $(\mathbf{X}_i, S_i)$, $i = 1, \ldots, N$, are assumed to be independent and identically distributed (i.i.d). Due to the i.i.d assumption, from here on, we ignore the subindex $_i$ whenever there is no confusion. The perfect PPV of the anchor variable, i.e., $\mathrm{PPV} = p(Y = 1|S = 1) = 1$, implies that Y is known to be 1 for those with $S = 1$. However Y can be either 0 or 1 when S is 0. Thus the anchor-positive patients, those with $(\mathbf{X}, S = 1)$, are automatically labeled as cases with $Y = 1$, and the anchor-

negative patients, those with $(\mathbf{X}, S = 0)$, are a mixture of cases and controls. The anchor variable also has a constant sensitivity, explicitly written as

$$p(S = 1|Y = 1, \mathbf{X}) = p(S = 1|Y = 1) = c,$$

where c is a constant between 0 and 1. That is, given $Y = 1$, S and $\mathbf{X}$ are independent. By Bayes rule, it is easy to show that

$$c = p(S = 1)/p(Y = 1). \tag{3.1}$$

Let $P(\mathbf{X}; \boldsymbol{\beta}) = p(Y = 1|\mathbf{X}; \boldsymbol{\beta})$ denote the phenotyping model of interest, which can be any practically reasonable parametric model to relate Y and $\mathbf{X}$. As an illustration, we consider

$$\text{logit } P(\mathbf{X}; \boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}. \tag{3.2}$$

Let q be the phenotype prevalence, i.e., $q = \int P(\mathbf{X}; \boldsymbol{\beta})dF(\mathbf{X})$, where $F(\mathbf{X})$ denotes the cumulative distribution function of $\mathbf{X}$.

3.3.2. Model calibration using positive-only EHR

For any risk prediction model, it is important to assess its calibration to ensure that it produces unbiased estimate of the phenotype risk. The classical Hosmer-lemeshow goodness-of-fit test (Hosmer, Lemeshow, and Klar, 1988) is widely applied if the case and control statuses are fully observed and if the risk prediction model is logistic. Specifically, the test can be summarized in three steps. First, it sorts all subjects by predicted risk. Second, it divides the sorted subjects into G risk subgroups. Third, it enumerates the number of observed and expected cases within each subgroup and calculate the statistic

$$T_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g p_g (1 - p_g)},$$

where O_g and E_g are the observed and expected number of cases in subgroup g, respectively, p_g is the predicted risk for the gth subgroup, and N_g is the number of subjects in subgroup g. The test statistic T_{HL} asymptotically follows a χ^2 distribution with $G - 2$ degrees of freedom.

Since the true phenotype status Y's are not observed for the unlabeled patients in the positive-

only data, the observed number of cases O_g cannot be obtained, hence the Hosmer-Lemeshow test becomes infeasible. Hereby we propose a new method that assesses calibration of a risk prediction model using positive-only data. We first divide the unlabeled patients into K pre-defined risk regions. Let $p\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\}$ denote the conditional probability of an unlabeled patient being a case in risk region $I_l, l = 1, ..., K$. Noting that $p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l|Y = 1, S = 0\} = p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l|S = 1\}$ because $\mathbf{X}$ and S are independent conditional on $Y = 1$ and $p(Y = 1 \mid S = 1) = 1$, we can rewrite $p\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\}$ as

$$\begin{aligned} & p\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\} \\ = & \frac{p(Y = 1, S = 0)p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l|Y = 1, S = 0\}}{p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\}} \\ = & \frac{p(Y = 1, S = 0)p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l|S = 1\}}{p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\}} \\ = & \frac{\{p(Y = 1) - p(S = 1)\}p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 1\}}{p(S = 1)p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\}}. \end{aligned}$$

Let $P(\mathbf{X};\widehat{\boldsymbol{\beta}})$ denote the predicted disease risk. We can therefore construct a model-free nonparametric estimator of the conditional probability $\widehat{p}^n\{Y = 1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S = 0\}$ as

$$\frac{\{q - N^{-1}\sum_{i=1}^N I(S_i = 1)\}[N^{-1}\sum_{i=1}^N I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) \in I_l, S_i = 1\}]}{\{N^{-1}\sum_{i=1}^N I(S_i = 1)\}[N^{-1}\sum_{i=1}^N I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) \in I_l, S_i = 0\}]}. \tag{3.3}$$

We further note that $p\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\}$ can be alternatively expressed as a function of parameters $\boldsymbol{\beta}$ and c:

$$\begin{aligned} p\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\} &= \frac{\int_{P(\mathbf{X};\boldsymbol{\beta})\in I_l} p(Y = 1, S = 0|\mathbf{X})dF(\mathbf{X})}{\int_{P(\mathbf{X};\boldsymbol{\beta})\in I_l} p(S = 0|\mathbf{X})dF(\mathbf{X})} \\ &\equiv \frac{\int_{P(\mathbf{X};\boldsymbol{\beta})\in I_l}(1 - c)P(\mathbf{X};\boldsymbol{\beta})dF(\mathbf{X})}{\int_{P(\mathbf{X};\boldsymbol{\beta})\in I_l}\{1 - cP(\mathbf{X};\boldsymbol{\beta})\}dF(\mathbf{X})}. \end{aligned}$$

Therefore a model-based parametric estimator of the same conditional probability can be obtained as

$$\widehat{p}^m\{Y = 1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S = 0\} = \frac{\sum_{i=1}^N(1 - \widehat{c})P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) \in I_l\}}{\sum_{i=1}^N\{1 - \widehat{c}P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})\}I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) \in I_l\}}, \tag{3.4}$$

where $\hat{c}$ is an estimate of c obtained as

$$\hat{c} = \widehat{p}(S_i = 1)/p(Y_i = 1) = N^{-1} \sum_{i=1}^{N} I(S_i = 1)/q$$

following (3.1). The parametric estimate $\widehat{p}^m\{Y = 1|P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}$ and nonparametric estimate $\widehat{p}^n\{Y = 1|P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}$ should be close to each other under the null hypothesis that the prediction model is correct. Therefore the extent to which these two estimates agree provides evidence for fit or lack-of-fit of the model.

Remark. *Note that although the expression $P(\mathbf{X}; \widehat{\boldsymbol{\beta}})$ appears in the expression (3.3), it does not have to be a true prediction model in order to yield a valid statistic $\widehat{p}^n\{Y = 1|P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}$. It is only used as a way to group the covariates. In fact, we could have chosen a completely arbitrary function to group the covariates, the statistic will still be valid. However, this is not the case in the construction of $\widehat{p}^m\{Y = 1|P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}$, which will not be a consistent estimator of $p\{Y = 1|P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \in I_l, S = 0\}$ if the prediction model $P(\mathbf{X}; \boldsymbol{\beta})$ is wrong. In this sense, we name the two estimators nonparametric and parametric respectively.*

When q is known

There are various situations in EHRs where the phenotype prevalence q is known. For example, it can be an educated guess by clinicians or obtained from other investigations. To assess the overall goodness-of-fit, we propose a novel statistic T^q. It aggregates the differences between the nonparametric and parametric estimates of the number of cases across $K - 1$ out of the K risk groups in the same spirit as the Hosmer-Lemeshow statistic, taking into account that the difference in the Kth interval is determined by the differences in the other $K - 1$ intervals. It additionally integrates the departure of the estimated phenotype prevalence, i.e., $\widehat{q} - q$, into the calibration assessment. Specifically, we construct the statistic T^q as

$$T^q = \mathbf{U}_K^{\mathrm{T}} \mathbf{V}_K^{-1} \mathbf{U}_K, \tag{3.5}$$

where $\mathbf{U}_K$ is a length K vector whose lth element ($l = 1, ..., K - 1$) is

$$U_{K,l} = \widehat{p}^n\{Y = 1|P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \in I_l, S = 0\} - \widehat{p}^m\{Y = 1|P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \in I_l, S = 0\},$$

and the Kth element is $U_{K,K} = \widehat{q} - q = N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) - q$. $\mathbf{V}_K$ is the asymptotic variance of $\mathbf{U}_K$. The test statistic T^q asymptotically follows a χ^2 distribution with K degrees of freedom under H_0, i.e. when the fitted model is correct. Detailed proofs and estimation of $\mathbf{V}_K$ are included in Appendix II.

When q is unknown

We of course have to estimate the phenotype prevalence q in (3.3) when it is unknown. To this end, we can form an estimator for q as $\widehat{q} = N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})$. Thus the non-parametric estimate $\widehat{p}^n\{Y=1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S=0\}$ becomes

$$\begin{aligned} & \widehat{p}^n\{Y=1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S=0\} \\ = \; & \frac{\{N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) - N^{-1}\sum_{i=1}^{N} I(S_i=1)\}[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) \in I_l, S_i=1\}]}{\{N^{-1}\sum_{i=1}^{N} I(S_i=1)\}[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) \in I_l, S_i=0\}]}. \end{aligned}$$

In the parametric estimation of $\widehat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S=0\}$ as shown in equation (3.4), c is estimated as $\hat{c} = \sum_{i=1}^{N} I(S_i=1)/\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})$. In this scenario we use the calibration statistic T^{nq} to aggregate the differences between $\widehat{p}^n\{Y=1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S=0\}$ and $\widehat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S=0\}$,

$$T^{nq} = \mathbf{U}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}^{-1}\mathbf{U}_{K-1}, \tag{3.6}$$

where $\mathbf{U}_{K-1}$ is a length $K-1$ vector whose lth element is

$$U_{K-1,l} = \widehat{p}^n\{Y=1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S=0\} - \widehat{p}^m\{Y=1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S=0\},$$

and $\mathbf{V}_{K-1}$ is the asymptotic variance of $\mathbf{U}_{K-1}$. Thus the test statistic T^{nq} asymptotically follows a χ^2 distribution with $K-1$ degrees of freedom when the fitted model is correct. Detailed proofs are included in Appendix III.

Remark. *Strictly speaking, the nonparametric estimator $\widehat{p}^n\{Y=1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S=0\}$ is not totally nonparametric when we replace the instance of p using $N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})$, in the sense that when the prediction model $P(\mathbf{X};\boldsymbol{\beta})$ is wrong, $\widehat{p}^n\{Y=1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S=0\}$ is no longer a consistent estimator of $p\{Y=1|P(\mathbf{X};\widehat{\boldsymbol{\beta}}) \in I_l, S=0\}$. Fortunately, we are only interested in the deviation between the parametric and nonparametric statistics, and they continue to be valid*

estimators of a same quantity under H_0 and deviate under alternative.

We emphasize that how the risk subgroups I_l, $l = 1, 2, \ldots, K$, are formed does not affect the development and validity of the proposed calibration methods. In the following, we form the risk subgroups according to the fitted probabilities $P(\mathbf{X}; \widehat{\boldsymbol{\beta}})$ in such a manner that the use of $K = 10$ risk subgroups results in cutpoints defined at the values $k/10, k = 1, 2, ..., 9$, and the risk subgroups contain all patients with fitted probability between the adjacent cutpoints, e.g., the first group contains patients with $P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \leqslant 0.1$, the second group contains patients with $0.1 < P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \leqslant 0.2$, $\ldots$, and the tenth group contains those with $0.9 < P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \leqslant 1$.

Estimation of calibration slope using positive-only EHR

The positive-only EHR can also be analyzed to estimate alternative calibration metrics, such as the calibration slope. Let $Z \equiv \text{logit} P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) = \mathbf{X}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}$ denote the logistically-transformed predicted risk $P(\mathbf{X}; \widehat{\boldsymbol{\beta}})$. The calibration slope is defined as the estimated log odds ratio parameter in a logistic regression model that relates the true phenotype status Y to Z,

$$\text{logit } P(Y = 1|Z; a, b) = a + bZ.$$

The calibration slope b should be equal to close to 1 if the fitted model is correct. Following Zhang et al., 2019, taking advantage of the two defining characteristics of anchor variable, the calibration intercept a, slope b, and anchor sensitivity c can be estimated simultaneously through maximizing the likelihood function of the observed data

$$\begin{aligned} L(a, b, c) &= \prod_{i=1}^{N} p(Z, S = 1)^S \times p(Z, S = 0)^{1-S} \\ &\propto \prod_{i=1}^{N} \{cP(Y = 1|Z; a, b)\}^S \times \{1 - cP(Y = 1|Z; a, b)\}^{1-S}. \end{aligned}$$

3.3.3. Predictive accuracy assessment among positive-only EHR

To evaluate predictive performance of a well calibrated model, we consider statistical measures including true positive rate (TPR$_v$), false positive rate (FPR$_v$), positive predictive value (PPV$_v$), negative predictive value (NPV$_v$) at a decision threshold v and area under the ROC curve (AUC). Given that anchor-positive patients are definitive cases by definition, we propose to assess model performance only in the unlabeled patients. We found that all these measures can be estimated

using the positive-only data because of the two defining characteristics of the anchor variable. To be specific, TPR_v is defined as $\text{TPR}_v \equiv p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 1, S = 0\}$. Recall that the anchor variable has the properties that S and $\mathbf{X}$ are independent given $Y = 1$ and there is no false positives, so $\text{TPR}_v = p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 1, S = 1\} = p\{P(\mathbf{X};\boldsymbol{\beta}) > v | S = 1\}$. This implies that TPR_v for the unlabeled can be estimated as $\widehat{\text{TPR}_v} = [N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S = 1\}]/[N^{-1}\sum_{i=1}^{N} I(S_i = 1)]$. Similarly, FPR_v is defined and further re-written as

$$\begin{aligned}
\text{FPR}_v &\equiv p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 0, S = 0\} \\
&= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\} - p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0, Y = 1\}}{p(Y = 0, S = 0)} \\
&= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\} - p\{P(\mathbf{X};\boldsymbol{\beta}) > v | S = 1\}\{q - p(S = 1)\}}{1 - q}
\end{aligned}$$

and can be estimated as

$$\begin{aligned}
\widehat{\text{FPR}_v} &= \frac{N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S_i = 0\}}{1 - N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})} \\
&- \frac{[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S_i = 1\}]\{N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) - N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}}{\{N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}\{1 - N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})\}}.
\end{aligned}$$

Similar estimates of PPV_v, NPV_v and AUC and their respective asymptotic properties are provided in Appendix IV.

3.4. Simulations

3.4.1. Simulation settings

We conducted extensive simulation studies to evaluate the validity and feasibility of the proposed Chi-squared statistics T^q and T^{nq} for model calibration among unlabeled patients, and reported their type I error rate and power at 0.05 nominal significance level. We also evaluated the consistency and efficiency of the proposed estimators for predictive accuracy measures using positive-only data.

In each Monte Carlo simulation, we firstly generated 9 independent risk predictors, with X_1, X_4 and X_7 from the normal distribution $N(5, 10)$, X_2, X_5 and X_8 from the Bernoulli distribution with success rate 0.5, and X_3, X_6 and X_9 from the logit transformed standard uniform distribution. We

then generated the binary outcome variable Y based on a logistic model of the form

$$\text{logit } p(Y = 1|\mathbf{X}; \boldsymbol{\beta}) = \beta_0 + \sum_{k=1}^{9} \beta_k X_k. \tag{3.7}$$

The regression coefficients were set at $(\beta_1, \beta_2, \beta_3) = (0.2, 0.4, 0.6)$, $(\beta_4, \beta_5, \beta_6) = (-1.0, -1.4, 1.8)$, $(\beta_7, \beta_8, \beta_9) = (-2.0, 2.4, 2.8)$, with (X_1, X_2, X_3), (X_4, X_5, X_6) and (X_7, X_8, X_9) resembling weak, moderate, and strong predictors respectively. The intercept parameter was selected in such a way that the phenotype prevalence was fixed at 20%. We fixed the anchor sensitivity c at 0.5, and generated the anchor variable S as follows: i) when $Y = 1$, S takes the value 1 with probability c, and ii) when $Y = 0$, S is always 0. For each Monte Carlo simulation, we ran 1000 replicates for each set-up.

3.4.2. Simulation results

As shown in Table 3.1, when the fitted model is correctly specified, the nonparametric and parametric estimates of cases were nearly identical across all risk groups. The quantile plots of the statistics T^q and T^{nq} suggest that the distributions of the proposed statistics were well approximated by χ^2 distributions with degrees of freedom K or $K-1$ when q is known or unknown respectively (Figures 3.1 and 3.2). As shown in Table 3.2, the type I error rates for both T^q and T^{nq} were very close to the 0.05 nominal level, independent of the number of groups K.

To evaluate the performance in terms of power, we deliberately fitted three mis-specified models: minor, which omitted weak predictors (X_1, X_2, X_3); moderate, which omitted moderate predictors (X_4, X_5, X_6); severe, which omitted both the weak and moderate predictors (X_1 through X_6) from the model in (3.7). Note that models can be mis-specified in many different ways, we hereby use these three misspecified models as illustration examples. The level of disagreement between nonparametrically and parametrically estimated number of cases increased as the misspecification became more severe (Table 3.1). The results in Table 3.3 suggest that the proposed calibration tests had low power in detecting weak misspecification but the power was substantially increased when the fitted model departed further from the true one. For example, at $K = 10$ and $N = 50,000$, the power was 0.12, 0.69 and 0.73 when q is known, and was 0.12, 0.59 and 0.63 when q is unknown, when the weak, moderate, or both weak and moderate predictors failed to be included in the fitted model respectively. The power also improved when the sample size increased as we expected. For example, when both weak and moderate predictors were missing from the model,

at $K = 10$, the power increased from 0.46 to 0.92 as the sample size increased from 30,000 to 80,000 when q was known, and increased from 0.38 to 0.87 when q was unknown (Table 3.3). As expected, knowing the phenotype prevalence q also helped increase power of the calibration test. For example, the power for detecting severe misspecification increased from 0.87 to 0.92 when $N = 80,000$ and $K = 10$. As illustrated in Table 3.3, the power of the proposed calibration tests T^q and T^{nq} seems to not rely very much on the selection of the number of groups K, as the observed power was very similar across $K = 5, 10, 12$. As shown in Table B.1, the calibration slopes estimated using positive-only data were very close to their respective benchmarks, which were calculated using fully observed Y, regardless the fitted model was correct or misspecified.

To evaluate consistency of the proposed estimators for predictive accuracy measures based on positive-only data, we compared these estimates to their benchmarks, which were calculated using the true phenotype status. As shown in Table 3.4, when the data-generating model (3.7) was fitted to each simulated dataset, the accuracy measure estimators, $\widehat{\mathrm{TPR}_v}$, $\widehat{\mathrm{FPR}_v}$, $\widehat{\mathrm{PPV}_v}$, $\widehat{\mathrm{NPV}_v}$ and $\widehat{\mathrm{AUC}}$ were very close to their respective benchmarks. The negligibly small biases indicated statistical consistency of the proposed estimators. To demonstrate the accuracy of the estimated asymptotic variance of these estimates, we compared their average asymptotic standard errors (ASE) and the empirical standard errors (ESE) based on 1000 replications. Some differences were observed with $N = 10,000$ as shown in Table B.2. However the differences became negligible when the sample size was increased to 80,000 (Table B.2). It appears that relatively large sample sizes are needed to reflect the asymptotic results for the proposed accuracy measures. We recommend using bootstrapping to estimate asymptotic standard errors in practice.

3.5. Data Example

3.5.1. Preliminary phenotyping models for primary aldosteronism

Primary aldosteronism (PA) is the most common cause of secondary hypertension, accounting for 5-10% of hypertensive patients (Oenolle et al., 1993; Rossi et al., 1996; Shigematsu et al., 1997). Once diagnosed, PA can be effectively treated with surgical or medical interventions. Unfortunately, this condition has been severely under-detected in general, so that the vast majority of affected patients (Mulatero et al., 2016) are left untreated. Consequently, developing EHR phenotyping models for PA will potentially increase detection rate. To do this, we have specified two case sets for PA. Based on each case set, we developed a corresponding preliminary phenotyping model

from Penn Medicine EHR data using the ML method (Zhang et al., 2019). One case set (S_1) was specified as whether the patient was included in an existing expert-curated PA research registry. Every patient included in this registry is a definitive PA case (Wachtel et al., 2016), because they underwent a diagnostic procedure, adrenal vein sampling, which was only performed on patients who were confirmed to have primary aldosteronism. We further supplemented the S_1-positive set by adding patients with laboratory test orders for adrenal vein sampling cortisol or aldosterone testing, which is only performed as part of the adrenal vein sampling procedure, and regarded them as S_2-positive. S_1 and S_2 will be used as the anchor variables.

The data set contained 6,319 patients who had orders for a PA screening laboratory test, with 149 (2.4%) cases labeled by S_1 and 196 (3.1%) labeled by S_2. Using S_1 or S_2 as the case set, we developed a corresponding logistic regression phenotyping model in the form of (3.2) with predictors including demographics, laboratory results, encounter metadata, diagnosis codes, selected variables that were available at the time of screening and some variables extracted from unstructured clinical text. The detailed forms of the phenotyping models are included in Supplementary Table A.6 & A.7 (Zhang et al., 2019). The two phenotyping models estimated the labeling sensitivity of S_1 and S_2 as 0.56 (95%CI: 0.46-0.65) and 0.61 (95%CI: 0.54-0.69), and PA prevalence as 4% (95%CI: 3%-5%) and 5% (95%CI: 4%-6%) respectively. In this study we focus on the calibration and prediction accuracy assessment of the developed models among the unlabeled patients.

3.5.2. Validation of PA phenotyping models

To validate the two positive-only EHR phenotyping models for PA, we firstly tested their calibration among the unlabeled patients and then evaluated their predictive performances. As shown in Table 3.5, we observed good agreement between the nonparametric and parametric estimates of PA cases for both sets. When the PA prevalence was unknown, the calibration statistic T^{nq} achieved p-values of 0.78 (T^{nq}=5.60) and 0.53 (T^{nq}=8.05) for S_1 and S_2 based models when the number of subgroups K was taken to be 10, indicating good calibration. When the PA prevalence in our population is known as 5% based on literature, then the calibration test T^q can be applied, which indicated excellent goodness-of-fit of the fitted models ($T^q = 5.84$ and 5.42; p-value$= 0.83$ and 0.86 for S_1 and S_2 based models respectively). We then evaluated their predictive accuracy using the proposed estimators. For both anchors, the model achieved $AUC > 0.98$. The PPV and TPR were estimated as 0.83 and 0.75 at the decision threshold 0.5 for S_1-based phenotyping model, and 0.81 and 0.79 for S_2-based model (Table 3.6). Charts for patients with $p(Y = 1|\mathbf{X}; \widehat{\boldsymbol{\beta}}) > 0.5$ were

reviewed by clinicians. According to the chart review results, the PPVs for models based on S_1 and S_2 was estimated at 0.78 and 0.77 respectively. Interestingly, the estimated PPVs based on the S_1 model were slightly higher than those based on the S_2 model, while the estimated TPRs were slightly lower. This might be due to the lower prevalence of PA cases among the patients unlabeled with S_2. The standard errors based on the S_2 model were generally smaller for the corresponding estimates. This is reasonable because S_2 labeled more cases.

To demonstrate the power of our proposed calibration tests using real world EHR data, for each anchor variable we also tested calibration of a baseline model which included only demographics (age, gender, race, ethnicity) and variables available at the time of screening (systolic and diastolic blood pressure, time between first office visit with blood pressure recorded to PA screening test, time between first clinical encounter to PA screening test), but did not include other important predictors (laboratory results, diagnosis codes, clinical notes, etc.). AUCs for the two baseline models were estimated at 0.79 and 0.78 respectively. However, the baseline models led to estimates of the phenotype prevalence $\widehat{q}$ to be unreasonably high (0.30 and 0.27 for S_1 and S_2 based models respectively). Both baseline models were ill calibrated as indicated by the extent of disagreement between nonparametric and parametric estimates of cases (Table 3.5). Fortunately, their lack of calibration were captured by both of the proposed calibration tests with p-values< 0.001 (T^q=228.6 and 232.6 assuming $q = 5\%$, T^{nq}=31.9 and 48.4 when q is unknown respectively) as shown in Table 3.5.

3.6. Discussion

Traditional methods for model validation depend on a fully labeled validation set. Our proposed methods pave the way to greatly alleviate the need for labor intensive manual labeling, because they do not require labeled controls and rely on the specification of an anchor variable to label the set of cases in positive-only data. Collectively, the anchor variable framework offers a highly efficient approach to EHR phenotyping as well as a cost effective data resource for model validation. As anchor variables continue to be developed for a wide variety of phenotypes, we expect that the methods for model validation proposed here, together with the maximum likelihood approach for model development proposed previously (Zhang et al., 2019), will not only advance accurate EHR phenotyping algorithm development but also facilitate their generalizability, since a well chosen anchor variable may be more easily transferred across institutions.

Unlike the Hosmer-Lemeshow test whose result depends markedly on the number of risk subgroups, our proposed calibration tests are robust to the number of subgroups. As with any goodness-of-fit tests, these tests do not necessarily work properly in the presence of sparse cells. This is also why we proposed grouping based on fixed values instead of percentiles of fitted probabilities, since the expected number of cases can be very low in the lower percentiles, especially when the phenotype prevalence is low. We recommend that the within-group expected number of cases be taken into consideration in the selection of the number of groups K. Similar as the Hosmer-Lemeshow goodness-of-fit test, the power of our proposed calibration test can be low when the sample size is not large. However, as observed in the analysis of the PA data (Table 3.5), even when the sample size is moderate, the difference between the nonparametric and parametric estimates of the number of cases is still informative of lack-of-fit, while the good agreement between the two estimates would lend strong support to the good calibration.

The proposed calibration statistics asymptotically follow a χ^2 distribution when the fitted model is correctly specified provided that the anchor variable is well defined, i.e., the anchor variable has perfect PPV and constant sensitivity. Thus with a validated anchor variable, significance of the proposed calibration tests informs goodness-of-fit. Noting that the calibration results can be affected by either model misspecification or invalid anchor variable specification, further research is needed for extensions to discriminate the two causes.

Table 3.1: Contingency table of nonparametric vs. parametric estimates of number of cases in intervals of estimated probabilities, estimation of anchor sensitivity c and phenotype prevalence q, calibration statistic and p-value of the fitted model when $K = 10$. Sample size $N = 50,000$. Correct: fitted model is correctly specified; Minor: fitted model missing weak predictors (X_1, X_2, X_3); Moderate: fitted model missing moderate predictors (X_4, X_5, X_6); Severe: fitted model missing weak and moderate predictors $(X_1 - X_6)$; Npara: nonparametric; Para: parametric.

	Correct		Minor		Moderate		Severe	
Interval	Npara	Para	Npara	Para	Npara	Para	Npara	Para
0.0_0.1	98	98	122	124	326	335	338	348
0.1_0.2	100	100	133	126	391	365	413	385
0.2_0.3	108	109	140	135	391	375	411	396
0.3_0.4	121	120	151	149	395	392	415	411
0.4_0.5	138	137	168	169	410	417	428	435
0.5_0.6	162	162	194	198	438	454	453	470
0.6_0.7	200	200	238	244	483	507	496	522
0.7_0.8	268	268	314	323	562	589	572	600
0.8_0.9	427	427	492	503	723	739	721	736
0.9_1.0	3483	3483	3198	3178	1331	1278	1239	1185
Estimation of (c, q) and their standard deviation								
$\widehat{c}$	0.500 (0.006)		0.498 (0.006)		0.484 (0.011)		0.482 (0.011)	
$\widehat{q}$	0.205 (0.002)		0.205 (0.003)		0.211 (0.005)		0.212 (0.005)	
Calibration results when q is known								
T^q	9.66		12.12		23.28		23.99	
p-value	0.47		0.28		0.01		0.01	
Calibration when q is unknown								
T^{nq}	8.64		10.88		19.56		20.21	
p-value	0.47		0.28		0.02		0.02	

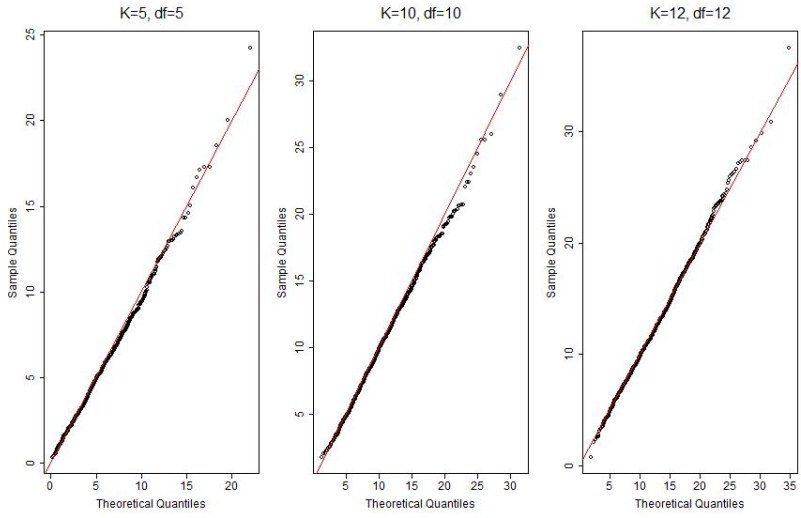

Figure 3.1: QQ plot of the calibration statistic T^q when $K = 5, 10, 12$ and sample size $N = 50,000$.

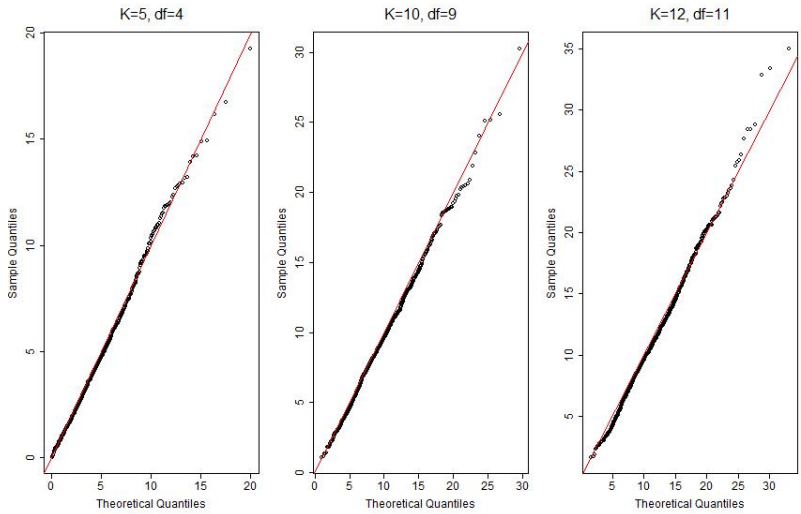

Figure 3.2: QQ plot of the calibration statistic T^{nq} when $K = 5, 10, 12$ and sample size $N = 50,000$.

Table 3.2: Type I error of the proposed calibration tests when we partition the estimated probabilities into $K = 5, 10, 12$ groups at different sample sizes.

	$K =$	5	10	12
T^q	$N = 30,000$	0.045	0.046	0.042
	$N = 50,000$	0.042	0.047	0.045
	$N = 80,000$	0.042	0.049	0.039
T^{nq}	$N = 30,000$	0.041	0.048	0.045
	$N = 50,000$	0.049	0.046	0.049
	$N = 80,000$	0.048	0.048	0.049

Table 3.3: Power of the proposed calibration test for detecting model misspecification at different numbers of subgroups K and sample size N. For model misspecification: Minor: fitted model missing weak predictors (X_1, X_2, X_3); Moderate: fitted model missing moderate predictors (X_4, X_5, X_6); Severe: fitted model missing weak and moderate predictors $(X_1 - X_6)$.

		Minor			Moderate			Severe		
	$K =$	5	10	12	5	10	12	5	10	12
T^q	$N = 30,000$	0.10	0.09	0.10	0.46	0.43	0.43	0.47	0.46	0.44
	$N = 50,000$	0.14	0.12	0.11	0.70	0.69	0.68	0.73	0.73	0.71
	$N = 80,000$	0.20	0.17	0.16	0.91	0.92	0.91	0.93	0.92	0.92
T^{nq}	$N = 30,000$	0.10	0.09	0.08	0.36	0.33	0.30	0.38	0.38	0.33
	$N = 50,000$	0.14	0.12	0.12	0.58	0.59	0.57	0.64	0.63	0.63
	$N = 80,000$	0.24	0.20	0.19	0.81	0.85	0.84	0.85	0.87	0.87

Table 3.4: Predictive accuracy measures evaluated among unlabeled patients when true phenotype status observed vs. estimated using positive-only data for $N = 10,000$ and $N = 80,000$. Results based on the mean over 1000 replications.

		Y observed (Benchmark)					Positive-only				
	Cutoff	PPV	NPV	TPR	FPR	AUC	$\widehat{\text{PPV}}$	$\widehat{\text{NPV}}$	$\widehat{\text{TPR}}$	$\widehat{\text{FPR}}$	$\widehat{\text{AUC}}$
$N = 10,000$	0.2	0.677	0.994	0.959	0.059	0.990	0.678	0.995	0.962	0.059	0.991
	0.4	0.789	0.989	0.913	0.032	-	0.792	0.989	0.918	0.031	-
	0.6	0.868	0.981	0.855	0.017	-	0.873	0.982	0.860	0.016	-
	0.8	0.935	0.971	0.766	0.007	-	0.939	0.971	0.769	0.006	-
$N = 80,000$	0.2	0.675	0.995	0.961	0.059	0.990	0.675	0.995	0.961	0.059	0.990
	0.4	0.789	0.989	0.916	0.032	-	0.789	0.989	0.916	0.031	-
	0.6	0.870	0.982	0.857	0.016	-	0.871	0.982	0.858	0.016	-
	0.8	0.938	0.971	0.765	0.007	-	0.938	0.971	0.766	0.006	-

Table 3.5: Contingency table of nonparametric vs. parametric estimates of PA cases in intervals of estimated probabilities, estimation of anchor sensitivity and phenotype prevalence, and calibration statistic and p-value for the final and baseline prediction models built upon two anchors of PA when q is assumed known as 5% or unknown respectively when the number of subgroups K was taken to be 10. Final: final phenotyping model; Baseline: baseline phenotyping model; Npara: nonparametric; Para: parametric.

	Anchor S_1				Anchor S_2			
	Final		Baseline		Final		Baseline	
Interval	Npara	Para	Npara	Para	Npara	Para	Npara	Para
0_0.1	12	11	51	39	7	7	52	44
0.1_0.2	8	6	111	121	4	4	132	139
0.2_0.3	3	6	137	161	6	3	160	156
0.3_0.4	7	5	154	160	1	3	97	135
0.4_0.5	7	6	154	138	5	3	109	95
0.5_0.6	8	9	103	114	3	5	103	76
0.6_0.7	0	5	86	74	3	3	40	37
0.7_0.8	8	10	69	43	5	5	6	17
0.8_0.9	10	13	26	17	6	7	17	7
0.9_1	106	99	129	154	85	84	143	155
Estimation of c and q								
$\widehat{c}(SE)$	0.56 (0.05)		0.10 (0.02)		0.61 (0.04)		0.15 (0.02)	
$\widehat{q}(SE)$	0.04 (0.004)		0.30 (0.05)		0.05 (0.004)		0.27 (0.04)	
Calibration when q is known								
T^q	5.84		228.63		5.42		232.56	
p-value	0.83		< 0.001		0.86		< 0.001	
Calibration when q is unknown								
T^{nq}	5.60		31.94		8.05		48.35	
p-value	0.78		< 0.001		0.53		< 0.001	

Table 3.6: Estimated accuracy measures and their associated empirical standard errors (in parentheses) for identifying PA patients using our proposed method for positive-only data. Results based on mean over 1000 bootstrap replicates.

	Anchor S_1					Anchor S_2				
cutoff	$\widehat{\text{PPV}}$	$\widehat{\text{NPV}}$	$\widehat{\text{TPR}}$	$\widehat{\text{FPR}}$	$\widehat{\text{AUC}}$	$\widehat{\text{PPV}}$	$\widehat{\text{NPV}}$	$\widehat{\text{TPR}}$	$\widehat{\text{FPR}}$	$\widehat{\text{AUC}}$
0.1	0.413	0.997	0.89	0.03	0.983	0.395	0.998	0.906	0.028	0.986
	(0.124)	(0.001)	(0.041)	(0.005)	(0.008)	(0.072)	(0.001)	(0.025)	(0.004)	(0.005)
0.2	0.559	0.997	0.85	0.016	-	0.533	0.997	0.872	0.015	-
	(0.128)	(0.001)	(0.054)	(0.003)	-	(0.077)	(0.001)	(0.03)	(0.002)	-
0.3	0.666	0.996	0.815	0.009	-	0.639	0.997	0.844	0.01	-
	(0.122)	(0.001)	(0.064)	(0.002)	-	(0.074)	(0.001)	(0.034)	(0.002)	-
0.4	0.755	0.995	0.782	0.006	-	0.732	0.996	0.817	0.006	-
	(0.11)	(0.001)	(0.071)	(0.001)	-	(0.071)	(0.001)	(0.039)	(0.001)	-
0.5	0.828	0.994	0.748	0.003	-	0.81	0.996	0.787	0.004	-
	(0.093)	(0.002)	(0.078)	(0.001)	-	(0.063)	(0.001)	(0.042)	(0.001)	-
0.6	0.882	0.993	0.707	0.002	-	0.871	0.995	0.753	0.002	-
	(0.078)	(0.002)	(0.087)	(0.001)	-	(0.056)	(0.001)	(0.048)	(0.001)	-
0.7	0.915	0.992	0.66	0.001	-	0.906	0.994	0.708	0.001	-
	(0.065)	(0.002)	(0.098)	(0.001)	-	(0.052)	(0.001)	(0.054)	(0.001)	-
0.8	0.922	0.991	0.601	0.001	-	0.919	0.993	0.649	0.001	-
	(0.058)	(0.002)	(0.107)	(0.001)	-	(0.047)	(0.001)	(0.06)	(0.001)	-
0.9	0.899	0.989	0.518	0.001	-	0.915	0.991	0.571	0.001	-
	(0.068)	(0.003)	(0.12)	(0.001)	-	(0.053)	(0.002)	(0.066)	(0.001)	-

CHAPTER 4

ELECTRONIC HEALTH RECORD PHENOTYPING WITH AUTOMATED VARIABLE SELECTION USING ANCHOR-POSITIVE AND UNLABELED PATIENTS

4.1. Abstract

Positive-only data, referring to data for a group of labeled cases and a large number of unlabeled patients, can often be accessed in electronic health records (EHRs) with minimal effort for labeling. A critical feature of such data is that the labeled cases are representative of all cases and can often be ascertained through anchor variables. We previously developed a maximum likelihood method for training logistic phenotyping models and nonparametric methods for assessing model calibration and predictive accuracy. These methods paved the way for positive-only data to become an attractive data source for developing and validating EHR phenotyping models. They cannot accommodate high dimensional predictors, however. Adapting existing methods for high-dimensional variable selection requires non-trivial efforts of developing specialized algorithms to accommodate absence of gold standard controls. In this study, we propose a two-stage approach to building high-dimensional phenotyping models. Automated variable selection is enabled through a proxy phenotype that is available for all patients using existing methods such as adaptive LASSO. Then the previously proposed maximum likelihood method for the positive-only data is applied to build the phenotyping model with the predicted probability of the proxy from the first stage included as a predictor. The proposed method had higher predictive accuracy compared to the same high-dimensional method applied to fully labeled datasets of realistic sample sizes in simulation studies. It also demonstrated good performance when applied to the real EHR data derived from the University of Pennsylvania Health System (UPHS), which achieved an area under the ROC curve (AUC) of 0.82.

4.2. Introduction

Electronic health records (EHRs) have become a valuable resource for clinical and health service research, which has fostered the development of EHR phenotyping algorithms. EHR-based phenotyping is to identify patients with certain phenotype based on their clinical features, both structured and unstructured, documented in EHRs (Elkan and Noto, 2008; Hong, Liao, and Cai, 2019; Pathak, Kho, and Denny, 2013; Shivade et al., 2013; Yu et al., 2016). Conventionally, the development of

a phenotyping algorithm involves two important steps: assigning gold-standard labels to a training set, and selecting informative features that strongly characterize the phenotype. To label patients with regard to presence (case) or absence (control) of the interest phenotype requires clinical experts to retrospectively review each patient's medical record, which is labor intensive and time consuming. The considerable efforts usually limit the achievable sample size of the training set, which typically ranges from 400 to 600 (Ananthakrishnan et al., 2013; Bejan et al., 2012; Carroll, Eyler, and Denny, 2011; Carroll et al., 2012; Kumar et al., 2014; Liao et al., 2015, 2010; Love, Cai, and Karlson, 2011; Xia et al., 2013; Yu et al., 2017). In addition, for most diseases, the clinical practice workflows are not symmetric with regards to cases and controls, with only specific patients actively screened based on clinical suspicion. So oftentimes it is relatively easy to obtain a group of cases, for example, taking advantage of the existence of disease registries, but prohibitive to identify a large number of gold-standard controls based on clinical practice data. As a result, for many phenotypes, the most readily accessible labels for a cohort of EHR data are a group of gold-standard cases and a large number of unlabeled patients whose phenotype statuses are unknown. Henceforth, we refer to the EHR data with such incomplete labels as "positive-only" data (Zhang et al., 2019).

To reduce dependency on the gold-standard labeled training set, in Chapter 2 we have proposed a maximum likelihood approach for developing phenotyping models by efficiently leveraging such "positive-only" data (Zhang et al., 2019). The method is built upon the assumption that the labeled cases in the positive-only data are representative of all cases in the population. We rely on the specification of an anchor variable to label the set of cases. An anchor variable is a binary variable that is highly informative of the true phenotype status, characterized by perfect positive predictive value (PPV) and independent sensitivity (Elkan and Noto, 2008; Halpern et al., 2014, 2016). Thus its positivity labels a random subset of all cases, while its negativity is nondeterministic of a patient's phenotype status. It does require domain expertise to specify an anchor variable for the interest phenotype, but should frequently require less upfront effort as compared to manual labeling.

Although the proposed approach spares the extensive human resource needed for annotating the training set, it is based on informative features manually selected by domain experts, thus is suitable for developing phenotyping models when the number of potential predictors is relatively small, and not scalable to massive EHR cohorts. Feature selection is one of the main steps and a primary key

to success in the development of phenotyping models. Manual variable selection is time consuming and not ideal for high dimensional data. Therefore, various techniques have been applied to reduce the dimensionality space, such as LASSO (least absolute shrinkage and selection operator) or adaptive LASSO (Tibshirani, 1996; Zou, 2006). Yet, these techniques rely on the availability of the gold-standard labeled training set, and its accuracy is oftentimes compromised by the limited sample size.

For many diseases, the screening decision (Yes or No) is available in EHRs, and is informative of the corresponding disease risk, since in practice, the screening tests are only prescribed to individuals who are suspected to be at risk for a particular disease, making it a potential risk factor for disease presence. Compared to the binary indicator of screening decision, the probability of being screened serves as a better predictor for disease risk. This is because for each patient the screening decision is made by his/her physician based on this particular physician's knowledge and evaluation of the patient's physical condition, including clinical presentation, medical history, family history, laboratory results etc.. Thus the screening decisions varies among physicians and across institutions. Fitting a risk prediction model for the screening decision using the tremendous amount of information contained in EHRs will potentially justify the variance and better identify the ones that should be screened by learning from so many physicians' decision making to patients with similar conditions (Wang et al., 2020).

In this Chapter, we extend the MLE approach proposed in Chapter 2 to a two-stage strategy in order to accommodate automated feature selection. In the first stage, taking the screening decision as the proxy phenotype, we apply dimension reduction techniques to perform automated variable selection and build a risk prediction model for screening. In the second stage, taking the predicted probability of screening as a predictor for the phenotype, along with a short list of expert-hand-picked risk factors, we apply the previously proposed maximum likelihood based anchor learning method to develop the phenotyping model. The resultant reduction in the need for manual labeling and manual variable selection would yield improved phenotyping efficiency and scalability.

In organizing the paper, we first describe the data structure of positive-only data in the framework of an anchor variable, and illustrate the proposed two-stage method in details. We then demonstrate its performance using positive-only data through extensive simulation studies, and compare it to the conventional LASSO and adaptive LASSO performed on an annotated training set. Lastly we

apply the proposed two-stage method to build a risk prediction models for primary aldosteronism using data derived from Penn Medicine EHRs (Zhang et al., 2019).

4.3. Methods

4.3.1. Positive-only EHR data

Let Y denote the binary label for the phenotype (1: case, 0: control), T denote the binary indicator for being screened or not (1: yes, 0: no), and S denote the binary anchor variable (1: positive, 0: negative). Let $\mathbf{X} = (x_1, x_2, ..., x_p)$ and $\mathbf{Z} = (z_1, z_2, ..., z_q)$ denote clinical features extracted from EHRs with dimensions p and q respectively, and $p \ll q$. Here $(\mathbf{X}, \mathbf{Z}, T, S, Y)$ are considered as random variables from which a cohort of EHR data is drawn. For the positive-only EHR data, a cohort sample of N patients that is randomly drawn from the EHRs, we only get to observe $(\mathbf{X}_i, \mathbf{Z}_i, T_i, S_i)$, $i = 1, ..., N$ but not the true phenotype Y, and we assume the data are identically distributed and independent of each other (i.i.d).

For a well chosen anchor variable, the perfect PPV implies that $Y = 1$ whenever $S = 1$, i.e., $p(Y = 1|S = 1) = 1$. But Y can take either value 0 or 1 when $S = 0$. The independent anchor sensitivity says that given $Y = 1$, S and $(\mathbf{X}, \mathbf{Z})$ are independent, which can be formalized as

$$p(S = 1|Y = 1, \mathbf{X}, \mathbf{Z}) = p(S = 1|Y = 1) \equiv c,$$

where c is a constant between 0 and 1.

We use a logistic regression working model to relate Y and its predictors $\mathbf{X}$ and $\mathbf{Z}$, which is commonly implemented in the EHR setting, although our method is applicable to any parametric model that is reasonable for modeling binary outcome variables:

$$\text{logit } p(Y = 1|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) = \beta_0 + \mathbf{X}\boldsymbol{\beta}_X + \mathbf{Z}\boldsymbol{\beta}_Z. \tag{4.1}$$

For notational simplicity, we use $\mathbf{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ to denote $p(Y = 1|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$. Our goal is to predict the probability of $Y = 1$ for any patient $(\mathbf{X}, \mathbf{Z})$, regardless screened or not. In other words, the availability of T is not required.

4.3.2. Two-stage method for learning with positive-only data

In this section we describe a two-stage method for predicting the probability of $Y = 1$ based on positive-only EHR data, that performs automated variable selection and takes the intermediate probability of being screened as a risk factor.

Stage I The screening decision is usually informative of a patient's phenotype status, since for most diseases, the screening test is only prescribed to individuals that are suspected to be at risk for the particular disease. The goal of the first stage is to find a function that maps the high dimensional clinical features $\mathbf{Z}$ to the screening test T, for example, LASSO (least absolute shrinkage and selection operator) (Tibshirani, 1996):

$$\widehat{\eta}_0, \widehat{\boldsymbol{\eta}}_Z = \min_{\eta_0, \boldsymbol{\eta}_Z} \{\sum_{i=1}^{N} T_i(\eta_0 + \sum_{j=1}^{q} \eta_j Z_{ij}) + log(1 + exp(\eta_0 + \sum_{j=1}^{q} \eta_j Z_{ij})) + \lambda \sum_{j=1}^{q} |\eta_j|\}, \tag{4.2}$$

with $\lambda \geqslant 0$ as the penalty parameter. Thus the probability of $T = 1$ for each patient can be obtained as

$$\widehat{p}_{T_i} \equiv \widehat{p}(T_i = 1|\mathbf{Z}_i) = \frac{exp(\widehat{\eta}_0 + \mathbf{Z}_i \widehat{\boldsymbol{\eta}}_Z)}{1 + exp(\widehat{\eta}_0 + \mathbf{Z}_i \widehat{\boldsymbol{\eta}}_Z)} \tag{4.3}$$

Stage II In this stage we propose to use the anchor-based maximum likelihood method for fitting the predictive model for the phenotype status Y

$$logit\ p(Y = 1|\mathbf{X}, \widehat{p}_T; \boldsymbol{\alpha}) = \alpha_0 + \mathbf{X}\boldsymbol{\alpha}_X + \alpha_T \widehat{p}_T, \tag{4.4}$$

which includes a short list of "significant" predictors $\mathbf{X}$ that are hand-picked by domain experts and the fitted probability of screening $\widehat{p}_T$ as risk factors. As shown in Zhang et al. (2019), taking advantage of the two defining characteristics of anchor variable, the regression coefficients $\boldsymbol{\alpha}$ and anchor sensitivity c can be estimated simultaneously through maximizing the likelihood function

$$\begin{aligned} L(\boldsymbol{\alpha}, c) &= \prod_{i=1}^{N} p(X_i, \widehat{p}_{T_i}, S_i = 1)^{S_i} \times p(X_i, \widehat{p}_{T_i}, S_i = 0)^{1-S_i} \\ &\propto \prod_{i=1}^{N} \{cP(Y_i = 1|X_i, \widehat{p}_{T_i}; \boldsymbol{\alpha})\}^{S_i} \times \{1 - cP(Y_i = 1|X_i, \widehat{p}_{T_i}; \boldsymbol{\alpha})\}^{1-S_i}. \end{aligned}$$

Henceforth, we use $\mathbf{P}(\mathbf{X}, \widehat{p}_T; \widehat{\boldsymbol{\alpha}})$ to denote the estimated probability of phenotype presence by the

proposed two-stage method.

4.3.3. Predictive accuracy assessment using positive-only data

To evaluate predictive performance of a well calibrated model, we consider statistical measures including true positive rate (TPR_v), positive predictive value (PPV_v) at a decision threshold v, area under the ROC curve (AUC), and Precision-Recall curve, which is a plot of PPV (y-axis) and TPR (x-axis) for different risk thresholds. As proposed by Zhang et al., 2019, all these measures can be estimated using positive-only EHR data taking advantage of the two unique characteristics of the anchor variable. Since anchor-positive patients are definitive cases by definition, we propose to assess model performance only in the unlabeled patients. For example, TPR_v is defined as $\text{TPR}_v \equiv p\{\mathbf{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v | Y = 1, S = 0\}$. Since anchor variable S labels a random subset of all cases, we have $\text{TPR}_v = p\{\mathbf{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v | Y = 1, S = 1\} = p\{\mathbf{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v | S = 1\}$. This implies that TPR_v for the model developed using our proposed two-stage method can be estimated as $\widehat{\text{TPR}_v} = [N^{-1} \sum_{i=1}^{N} I\{\mathbf{P}(\mathbf{X}_i, \widehat{p}_{T_i}; \widehat{\boldsymbol{\alpha}}) > v, S = 1\}] / [N^{-1} \sum_{i=1}^{N} I(S_i = 1)]$. Similarly, PPV can is defined as

$$\begin{aligned} PPV_v &= p\{Y = 1 | P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v, S = 0\} \\ &= \frac{p\{P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v | Y = 1, S = 0\} p\{Y = 1, S = 0\}}{p\{P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v, S = 0\}} \\ &= \frac{p\{P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v | S = 1\} [p\{Y = 1\} - p\{Y = 1, S = 1\}]}{p\{P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) > v, S = 0\}}, \end{aligned}$$

and can be further estimated as

$$\widehat{PPV_v} = \frac{[N^{-1} \sum_{i=1}^{N} I\{\mathbf{P}(\mathbf{X}_i, \widehat{p}_{T_i}; \widehat{\boldsymbol{\alpha}}) > v, S_i = 1\}] \{N^{-1} \sum_{i=1}^{N} \mathbf{P}(\mathbf{X}_i, \widehat{p}_{T_i}; \widehat{\boldsymbol{\alpha}}) - N^{-1} \sum_{i=1}^{N} I(S_i = 1)\}}{\{N^{-1} \sum_{i=1}^{N} I(S_i = 1)\} [N^{-1} \sum_{i=1}^{N} I\{\mathbf{P}(\mathbf{X}_i, \widehat{p}_{T_i}; \widehat{\boldsymbol{\alpha}}) > v, S_i = 0\}]}$$

Details on the derivation and large sample properties of these accuracy measures can be found in (Zhang et al., 2020).

4.4. Simulation Studies

We carried out extensive simulation studies to evaluate the performance of the proposed two-stage phenotyping method using positive-only data, referred to as *Two-stage-LASSO* or *Two-stage-Adaptive-LASSO* when LASSO or adaptive LASSO was performed in the first stage respectively. We also included results from *One-step-LASSO* or *One-step-Adaptive-LASSO* as comparison benchmarks, where a set of either 400, 600 or 800 fully labeled data was used to fit standard LASSO or Adaptive LASSO with the true labels Y as the outcome variable.

4.4.1. Simulation Settings

To make the simulation more transferable into the real world, we generated model predictors by randomly selecting 326 features from UPHS EHR data, and bootstrapping the existing 10385 observations with replacement to achieve a sample size of 30,000. Details of the UPHS EHR data are described in Section 4.5.1. We denoted the first three features as $\mathbf{X}$, and the rest 323 as $\mathbf{Z}$. We then generated the binary indicator for screening test T according to the following model

$$logit\ p(T = 1|\mathbf{Z}, \eta_0, \boldsymbol{\eta}_Z) = \eta_0 + \mathbf{Z}\boldsymbol{\eta}_Z, \tag{4.5}$$

where we randomly selected 10 entries of $\boldsymbol{\eta}_Z$ to values $(0.2, 0.4, 0.6, 1.2, 1.4, 1.6, 2.2, 2.2, 2.6, 2.8)$ mimicking weak to strong relations, and fixed all other entries at zero. The intercept coefficient η_0 was chosen to fix the prevalence of T at 20%. We then generated the outcome variable Y based on the model

$$logit\ p(Y = 1|\mathbf{X}, \mathbf{Z}, \beta_0, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Z) = \beta_0 + \mathbf{X}\boldsymbol{\beta}_X + \mathbf{Z}\boldsymbol{\beta}_Z, \tag{4.6}$$

where $\boldsymbol{\beta}_X = (log(2.5), log(2.0), log(1.5))^\top$, $\boldsymbol{\beta}_Z = 1.25 \times \boldsymbol{\eta}_Z$, and β_0 was chosen to fix the prevalence of Y at 10%. The anchor variable S was then generated with anchor sensitivity c fixed at 0.5. For each $Y = 1$, the anchor variable S was generated according to a Bernoulli distribution with success rate c. For each $Y = 0$, S was always set to 0. In each Monte Carlo simulation, we randomly drew a sample of size 20,000 as the training set, holding the rest as the testing set. To implement the proposed two-stage method, Y was assumed unobserved in the training set. To implement the *Ideal learning*, we randomly drew $n_v = 400$, 600 or 800 samples within the 20,000 training data as the validation set in which Y was assumed observed. To compare the prediction performance of the two methods, we use AUC evaluated within the testing sample as the performance metric. All simulations results were based on 1000 replicates, and standard errors for AUCs were estimated empirically based on 1000 replicates.

4.4.2. Simulation Results

As shown in Table 4.1 and Figure 4.1, our proposed two-stage method achieved an AUC of 0.936 (SE: 0.004) when LASSO was used in the first stage, which was much higher than that of *One-step-Lasso* performed within an annotated validation set: 0.884 (SE: 0.019), 0.903 (SE: 0.013), and 0.912 (SE: 0.009) for validation set of size 400, 600, and 800 respectively. While the performance of *One-step-Lasso* depend on the size of validation set, for example, its AUC increased from 0.884

to 0.912 from $n_v = 400$ to $n_v = 800$, our proposed *Two-stage-Lasso* method does not require validation set. Yet it outperformed the *One-step-Lasso*'s AUC by 3% even when the sample size of validation set was increased to 800. In addition, the standard errors of its AUCs were much smaller than that of *One-step-Lasso*, indicating the proposed *Two-stage-Lasso* yielded more accurate and stable predictive performance. The Precision-Recall curve (Figure 4.2), which is a plot of PPV (y-axis) and TPR (x-axis) for different risk thresholds, also showed that *Two-stage-Lasso* achieved better prediction performance than *One-step-Lasso* even with a large validation set, $n_v = 800$. The results are similar when Adaptive LASSO was performed.

As shown in Table 4.2, the proposed *Two-stage-Lasso* correctly picked all 10 truly associated predictors $\mathbf{Z}$. On the contrary, this number was limited to 5, 6, and 6 when *One-step-Lasso* was applied in a validation set of size 400, 600, and 800 respectively. Table 4.3 shows the bias, asymptotic and empirical standard errors of estimated log-odds-ratio (logOR) for predictors $\mathbf{X}$ using proposed two-stage method when LASSO or Adaptive-Lasso was used in the first stage respectively. As demonstrated in Table 4.3, the two-stage method achieved consistent estimates of parameters $\boldsymbol{\beta}_X$, indicated by the biases being close to zeros. It also achieved consistent estimate of anchor sensitivity c and phenotype prevalence q as shown in Table 4.3.

4.5. Real Data Analysis

4.5.1. Application to the UPHS EHR data

PA is the most common cause of secondary hypertension, affecting 5% of hypertensive patients (Oenolle et al., 1993; Rossi et al., 1996; Shigematsu et al., 1997). It can be treated well by unilateral adrenalectomy or targeted medications. Unfortunately, PA is not recognized in many affected patients (Mulatero et al., 2016), so methods for EHR phenotyping have the potential to dramatically improve care for these patients. 10385 Penn Medicine primary care hypertensive patients who had at least two encounters between January 01, 1997 and July 01, 2017 were extracted. 6317 of them had an order for a PA screening laboratory test, the ratio of aldosterone to renin in the blood. 326 clinical features were used as potential predictors including demographics, encounter information, blood pressures, laboratory tests, diagnosis codes, and clinical notes. Because lab test ordering is non-random and potentially informative of phenotype status, we created binary variables indicative of result presence and included these binary indicators as well as the original lab test variables, with missing values replaced by zero. Missing values in variables other than laboratory results were

imputed with respective median values. In addition, variables with highly positively skewed distributions were log-transformed, and all continuous variables were standardized. Since the focus of this work is on demonstrating the performance of the proposed two-stage method with well-chosen set of cases, we leveraged an existing expert-curated PA research registry (Wachtel et al., 2016) in which every patient was a definitive PA case (S_1-positive). The patients were identified for the registry because they underwent a diagnostic procedure, adrenal vein sampling, which in our practice is only performed on patients definitively diagnosed with primary aldosteronism. We further S_1-positive set by adding patients with laboratory test orders for adrenal vein sampling cortisol or aldosterone testing, which is only performed as part of the adrenal vein sampling procedure, and regarded them as S_2-positive. Of all patients, 456 (4.4%) were S_1-positive and 569 (5.5%) were S_2-positive. For each anchor S_1 and S_2, we developed a corresponding preliminary phenotyping model for PA using the proposed two-stage method.

To apply the proposed two-stage method, in the first stage we used the penalized logistic regression, LASSO and Adaptive LASSO, to train the predictive model for the screening test. Using the PA screening test as outcome variable, all 326 clinical features as predictors, LASSO and Adaptive LASSO performed both variable selection and regularization. LASSO and Adaptive LASSO reduced the first stage candidate predictors to 82 and 130 respectively, and output estimated probability of being screened for the next step. The top 20 features for the stage I model are shown in Supplementary Table C.2. In the second stage, we use the proposed likelihood approach as described above to fit a standard logistic regression using the estimated probability of being screened and a short list of predictors identified by domain experts (fitted model shown in Supplementary Table C.3). To evaluate the predictive performance of the developed model, we used consistent estimators for TPR, PPV, and AUC based on positive-only data as described above in Section 4.3.3, and estimated them via 10-fold cross-validation.

When LASSO was performed in the first stage, using anchor S_1, the proposed two-stage method estimated the label sensitivity c as 0.68 (SE: 0.07), and consequently estimated the PA prevalence as 6% (SE: 0.7%). It achieved AUC of 0.82 (Figure 4.3), TPR and PPV of 0.33 and 0.31 respectively at threshold 0.3 (Table 4.4, Figure 4.4). When S_2 was used, the label sensitivity was estimated as 0.78 (SE: 0.06), and PA prevalence was estimated at 7% (SE: 0.6%). AUC was estimated as 0.82 (Figure 4.3), and at threshold 0.3, TPR and PPV were estimated as 0.3 and 0.23 respectively (Table

4.4, Figure 4.4). The results were similar when Adaptive LASSO was performed in the first stage as shown in Supplementary Table C.4.

4.6. Discussion

A key step to the development of EHR phenotyping models is the accurate selection of informative features that strongly characterize the phenotype. To do this, conventional methods either rely on manual variable selection or dimensionality reduction techniques. Manual selection takes significant time and expert knowledge, and is not scalable to high throughput data. Existing dimensionality reduction techniques oftentimes depends on a large number of gold-standard labeled cases and controls, while the labeling process is labor intensive and time consuming. Our proposed method uses the screening test as an easy surrogate to do variable selection, which builds a robust model for it leveraging the large number of EHR data, and then takes the fitted probability as a predictor and develops a phenotyping model based on a random sample of cases and a large number of unlabeled patients. This method leverages a domain-expertise-summarizing variable, anchor variable, to label the set of cases, which requires modest upfront effort from domain experts compared to manual labeling. Compared to standard strategies, the proposed method dramatically decreases the need for labor-intensive chart annotation and feature selection.

Our proposed two-stage method does not require gold-standard controls. Yet it outperformed the one-step LASSO and Adaptive LASSO for which a set of gold-standard labeled cases and controls were necessary on detecting individuals at-risk for disease. The surpass persists even when the number of labeled patients was increased to 800, which was usually infeasibly high in practice. In addition, our proposed method left a lot of flexibility in the first stage, where any supervised learning algorithms can be used as long as it performs dimensionality reduction and develops a robust model for screening test and outputs the fitted probability for the second stage, such as LASSO, adaptive LASSO, Post LASSO, Ridge regression etc.

Considering the large sample size of EHR data, our proposed method is expected to accurately extract important features and develop a robust model for screening test in the first stage. These intermediate products can be used to instruct physician's screening decisions, consequently facilitate early diagnosis of disease, and ultimately improve patients healthcare outcomes. They can be particularly beneficial when the screening tests are expensive or invasive, for example, genetic testing.

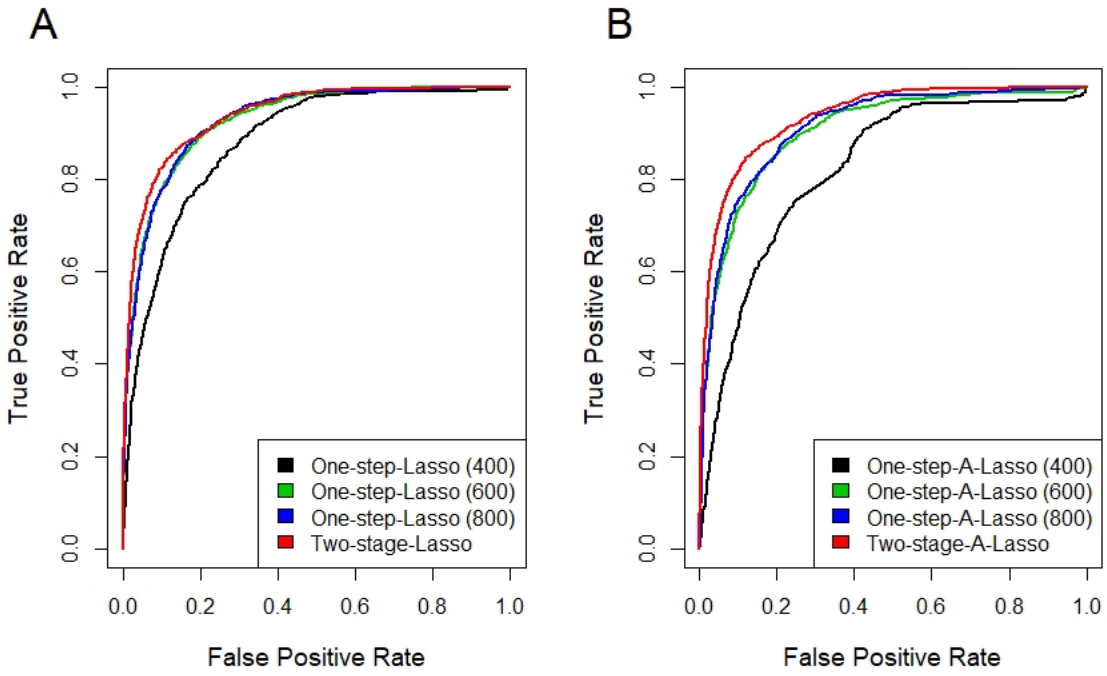

Figure 4.1: ROC plot for testing set when LASSO (A) or Adaptive-LASSO (B) was performed in the validation set of size 400, 600, 800 (One-step) or in the first stage of the proposed two-stage method (Two-stage).

Table 4.1: Estimated mean and empirical standard error of AUC when LASSO or Adaptive LASSO was applied in the first stage respectively.

	LASSO		Adaptive LASSO	
Validation set	One-step	Two-stage	One-step	Two-stage
$n = 400$	0.884 (0.019)	0.936 (0.004)	0.696 (0.166)	0.935 (0.004)
$n = 600$	0.903 (0.013)	-	0.798 (0.143)	-
$n = 800$	0.912 (0.009)	-	0.85 (0.101)	-

Table 4.2: Number of true predictors captured by when LASSO or Adaptive LASSO was applied in validation set or the first stage of the two-stage method respectively.

	LASSO		Adaptive LASSO	
Validation set	One-step	Two-stage	One-step	Two-stage
$n = 400$	5	10	1	8
$n = 600$	6	-	3	-
$n = 800$	6	-	4	-

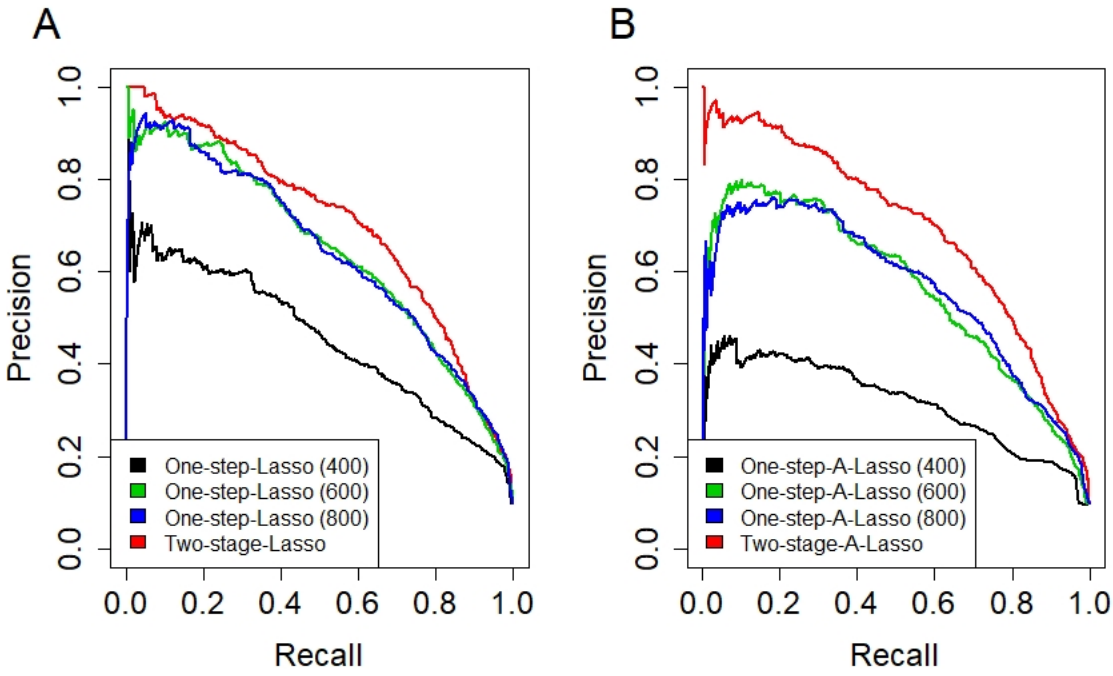

Figure 4.2: Utility plot for testing set when LASSO (A) or Adaptive-LASSO (B) was performed in the validation set of size 400, 600, 800 (One-step) or in the first stage of the proposed two-stage method (Two-stage).

Table 4.3: Estimated mean, asymptotic standard error (ASE), and empirical standard error (ESE) of logOR for $\mathbf{X}$, anchor sensitivity (c), and phenotype prevalence (q) using the two-stage method when LASSO or Adaptive LASSO was applied in the first stage respectively.

		LASSO			Adaptive LASSO		
	True value	Bias	ASE	ESE	Bias	ASE	ESE
$\widehat{\alpha}_{x_1}$	0.916	-0.016	0.149	0.153	0.003	0.149	0.153
$\widehat{\alpha}_{x_2}$	0.693	0.017	0.112	0.118	0.027	0.112	0.118
$\widehat{\alpha}_{x_3}$	0.405	0.007	0.146	0.150	0.009	0.147	0.151
$\widehat{c}$	0.500	-0.027	0.027	0.029	-0.031	0.027	0.030
$\widehat{q}$	0.100	0.006	0.006	0.006	0.007	0.006	0.006

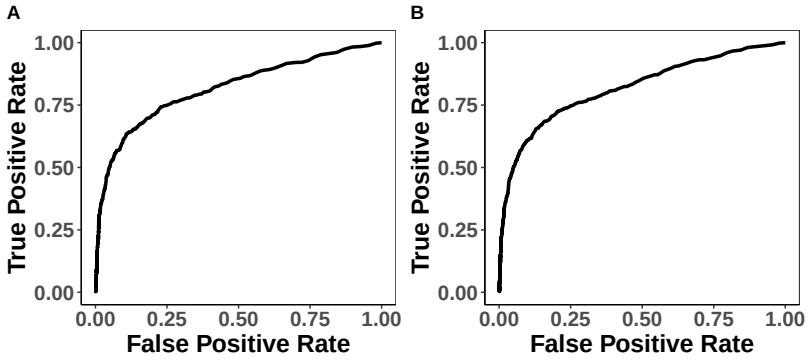

Figure 4.3: ROC plot for the positive-only EHR data in UPHS study when using anchor S_1 (A) and S_2 (B) while applying LASSO in the first stage.

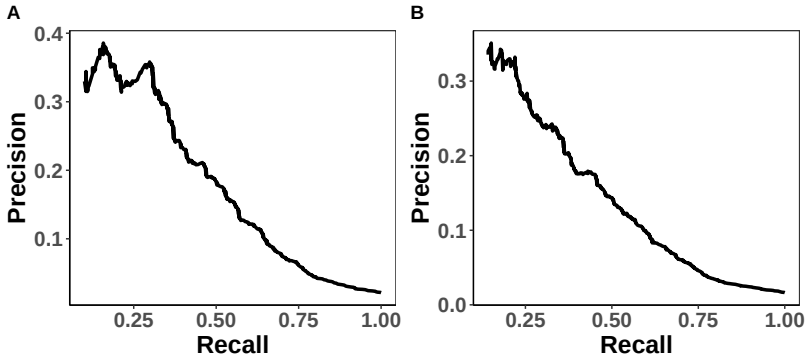

Figure 4.4: Utility plot for the positive-only EHR data in UPHS PA study when using anchor S_1 (A) and S_2 (B) while applying LASSO in the first stage.

Table 4.4: Estimated accuracy measures for the PA phenotyping model derived from positive-only EHR data using proposed two-stage method when LASSO was performed in the first stage.

	Anchor S_1		Anchor S_2	
Threshold	$\widehat{TPR}$	$\widehat{PPV}$	$\widehat{TPR}$	$\widehat{PPV}$
0.050	0.779	0.050	0.786	0.036
0.100	0.673	0.085	0.691	0.062
0.150	0.572	0.129	0.608	0.094
0.200	0.509	0.177	0.524	0.130
0.250	0.397	0.234	0.439	0.178
0.300	0.327	0.310	0.343	0.234
0.350	0.235	0.326	0.265	0.261
0.400	0.164	0.376	0.207	0.320
0.450	0.092	0.343	0.139	0.353
0.500	0.064	0.338	0.093	0.355

CHAPTER 5

DISCUSSION

In this dissertation, we have developed novel methods for the development and validation of accurate EHR phenotyping models that do not require labeled controls. Our proposed methods build upon a random sample of positive cases and a large number of unlabeled patients. And we leverage domain expertise summarized in the form of an anchor variable to identify the random sample of cases. In Chapter 2, using a logistic regression model to relate the binary phenotype and its predictors, we first proposed a maximum likelihood approach to developing an accurate risk prediction model within the anchor variable framework, which outperformed standard strategies. We then applied the proposed method to develop phenotyping models for primary aldosteronism using data extracted from Penn Medicine EHRs. In Chapter 3, motivated by the lack of methods for validating model performance using positive-only data, we derived model-free estimate for the number of cases in risk subgroups among unlabeled patients, and subsequently proposed a novel calibration test aggregating differences between model-free and model based estimated numbers of cases across risk subgroups. We also developed consistent estimators for predictive accuracy measures including TPR, FPR, PPV, NPV, and AUC, and derived their asymptotic properties. We applied the proposed validation methods to assess calibration and prediction accuracy of the PA phenotyping models developed in Chapter 2. In Chapter 4, to decrease the need for manual variable selection, we extended the likelihood approach proposed in Chapter 2 to a two-stage strategy, where a risk prediction model for screening test was built in the first stage, and the maximum likelihood approach was applied to build a logistic regression phenotyping model taking the fitted probability of being screened as a risk predictor. We demonstrated that our method outperformed conventional algorithms through simulation studies and application to real-world EHRs.

Since the specification of anchor variables requires modest upfront effort from clinical expertise, our proposed methods dramatically decreases the need for labor-intensive chart annotation and prospective phenotyping. Moreover, the anchor concept may itself be more easily transferred rather than the full model, thus our proposed methods will facilitate phenotype model development and transferability for a wide variety of EHR clinical decision support and research applications.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Appendix I: Proof of Parameter Identifiability

The joint pdf of $(\mathbf{X}, S)$ can be derived as

$$f(\mathbf{X}) \left\{ \frac{hP(\mathbf{X};\boldsymbol{\beta})}{\int P(\mathbf{X};\boldsymbol{\beta}) f(\mathbf{X}) d\mathbf{X}} \right\}^{I(S=1)} \left\{ 1 - \frac{hP(\mathbf{X};\boldsymbol{\beta})}{\int P(\mathbf{X};\boldsymbol{\beta}) f(\mathbf{X}) d\mathbf{X}} \right\}^{I(S=0)}$$

with unknown parameters $h, \boldsymbol{\beta}, f(\mathbf{X})$. Note that we have a random sample of $(\mathbf{X}, S)$'s, hence the observations $\mathbf{X}_i$'s form a random sample of $f(\mathbf{X})$. Thus, obviously, $f(\mathbf{X})$ is identifiable. Likewise, the observations S_i's form a random sample of the pmf of S, hence h is obviously identifiable. Assume that $\boldsymbol{\beta}$ is not identifiable. Then there exists $\boldsymbol{\beta}^*$ so that $p(\mathbf{X}, S; \boldsymbol{\beta}) = p(\mathbf{X}, S; \boldsymbol{\beta}^*)$ for all $(\mathbf{X}, S)$ pairs, which can be formalized as:

$$\begin{aligned} & \left\{ \frac{hP(\mathbf{X};\boldsymbol{\beta})}{\int P(\mathbf{X};\boldsymbol{\beta}) f(\mathbf{X}) d\mathbf{X}} \right\}^{I(S=1)} \left\{ 1 - \frac{hP(\mathbf{X};\boldsymbol{\beta})}{\int P(\mathbf{X};\boldsymbol{\beta}) f(\mathbf{X}) d\mathbf{X}} \right\}^{I(S=0)} \\ = & \left\{ \frac{hP(\mathbf{X};\boldsymbol{\beta}^*)}{\int P(\mathbf{X};\boldsymbol{\beta}^*) f(\mathbf{X}) d\mathbf{X}} \right\}^{I(S=1)} \left\{ 1 - \frac{hP(\mathbf{X};\boldsymbol{\beta}^*)}{\int P(\mathbf{X};\boldsymbol{\beta}^*) f(\mathbf{X}) d\mathbf{X}} \right\}^{I(S=0)}. \end{aligned}$$

Letting $S = 1$, this equation reduces to

$$\frac{hP(\mathbf{X};\boldsymbol{\beta})}{\int P(\mathbf{X};\boldsymbol{\beta}) f(\mathbf{X}) d\mathbf{X}} = \frac{hP(\mathbf{X};\boldsymbol{\beta}^*)}{\int P(\mathbf{X};\boldsymbol{\beta}^*) f(\mathbf{X}) d\mathbf{X}},$$

or equivalently

$$\frac{P(\mathbf{X};\boldsymbol{\beta}^*)}{P(\mathbf{X};\boldsymbol{\beta})} = \frac{\int P(\mathbf{X};\boldsymbol{\beta}^*) f(\mathbf{X}) d\mathbf{X}}{\int P(\mathbf{X};\boldsymbol{\beta}) f(\mathbf{X}) d\mathbf{X}}. \tag{A.1}$$

Note that the right hand side of (A.1) is a positive constant, hence the left hand side is also a constant. When the covariates $X_j \to \infty$ if $\beta_j > 0$ and $X_j \to -\infty$ if $\beta_j \leq 0$, regardless what the intercept term is, $P(\mathbf{X};\boldsymbol{\beta}) \to 1$ and $P(\mathbf{X};\boldsymbol{\beta}^*)$ converges to either 1 or 0. Thus, the left hand side goes to either 1 or 0. Because the right hand side is positive, hence we further conclude that the left hand side goes to 1. Because it has to be a constant, this implies that the left hand side is always 1, i.e. $P(\mathbf{X};\boldsymbol{\beta}^*) = P(\mathbf{X};\boldsymbol{\beta})$ at all $\mathbf{X}$. Thus $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, i.e. $\boldsymbol{\beta}$ is identifiable. Once $\boldsymbol{\beta}, h, f(\mathbf{X})$ are

indentifiable, the disease prevalence $q = \int P(\mathbf{X}; \boldsymbol{\beta}) f(\mathbf{X}) d\mathbf{X}$ is identifiable by definition.

Appendix II: Phenotyping with an anchor of varying sensitivity

Suppose that the population can be divided into K strata of size N_k, where the k^{th} stratum is indicated by $z_k,\ k = 1, ..., K$. Within each substratum, the anchor has a constant sensitivity. For notational convenience, Z is included as a component of predictors $\mathbf{X}$. The problem is then formalized as stratified conditional independence,

$$p(S = 1|Y = 1, \mathbf{X}) = p(S = 1|Y = 1, Z = z_k) = c_k, k = 1, ..., K.$$

The logistic model (2) remains the working model for prediction, where Z, if not predictive, has a log odds ratio equal to zero. The probability of observed data of an anchor-positive case and an unlabeled patient in each stratum is generalized as below to reflect variation in anchor sensitivity,

$$\begin{aligned} p(\mathbf{X}, S = 1) &= c_k p(Y = 1|\mathbf{X}) p(\mathbf{X}), \\ p(\mathbf{X}, S = 0) &= \{1 - c_k p(Y = 1|\mathbf{X})\} p(\mathbf{X}). \end{aligned}$$

The likelihood function becomes the probability of observed data across all K strata,

$$l(\boldsymbol{\beta}, \{c_k\}) \propto \sum_{i=1}^{N} \sum_{k=1}^{K} I(Z_i = z_k)[S_i \text{log}\{c_k P(\mathbf{X}_i; \boldsymbol{\beta})\} + (1 - S_i) \text{log}\{1 - c_k P(\mathbf{X}_i; \boldsymbol{\beta})\}]$$

Parameter identifiability can be shown similarly as in the situation of constant anchor sensitivity. The MLE of $\boldsymbol{\beta}$ and $\{c_k,\ k = 1, ..., K\}$ can be obtained by maximizing the log likelihood function $l(\boldsymbol{\beta}, \{c_k\})$. Let $q_k \equiv p(Y = 1|z_k)$ denote the phenotype prevalence in the k^{th} stratum. Then q_k can be estimated as $\hat{h}_k/\hat{c}_k$, where $\hat{h}_k$ is the estimated anchor prevalence in the k^{th} stratum, $\hat{p}(S = 1|Z = z_k) = \sum_{i=1}^{N} I(Z_i = z_k, S_i = 1)/N_k$. The overall phenotype prevalence q can then be estimated as a weighted summation of q_k's, $\hat{q} = \sum_{k=1}^{K} N_k \hat{q}_k / N$.

Appendix III: Simulation Results

As shown in Table A.1, the proposed maximum likelihood approach achieved consistent and efficient estimate of the odds ratio parameters $\boldsymbol{\beta}$, the anchor sensitivity c, and the phenotype prevalence q at different phenotype prevalence set-ups (5%, 10%, 15%, 20%).

To investigate the variation in model performance with respect to anchor sensitivity c, we considered

two values for c, 0.5 and 0.2. With 10,000 observations in the training set and phenotype prevalence 10%, taking 0.5 as the decision threshold, the anchor variable helped identify 536 cases when c equaled 0.5 and 797 cases when c equaled 0.2. The estimates of c and q appeared to be consistent in both cases, although those at $c = 0.2$ had wider confidence intervals as expected (Table A.2). The OR estimates $\hat{\boldsymbol{\beta}}$ appeared somewhat biased when $c = 0.2$. However, the bias largely disappeared when the size of the training set was increased to $20,000$ (Table A.2).

To assess the model performance when the anchor sensitivity varies, we simulated a new population following the same steps as above, except that it consisted of two strata indicated by a binary variable Z that takes value 0 or 1 with probability $p(Z = 1) = 0.5$. Conditional on the phenotype the anchor variable was independent of all predictors within each strata, with $c_1 = p(S = 1|Y = 1, Z_1) = 0.2$, and $c_2 = p(S = 1|Y = 1, Z_2) = 0.8$. Applying the proposed method that takes the stratified conditional independence into account, the parameter estimates, $\hat{\boldsymbol{\beta}}$, $\hat{c_1}$, $\hat{c_2}$, $\hat{q}$, appeared consistent (Table A.3). The anchor sensitivity c_1 was estimated to be 0.201 (SE: 0.025; 95% CI: 0.151, 0.251), c_2 was estimated to be 0.800 (SE: 0.022; 95% CI: 0.757, 0.843), and consequently the phenotype prevalence q was estimated to be 0.100 (SE: 0.004; 95% CI: 0.091, 0.109). On the other hand, if the variation in anchor sensitivity failed to be recognized, the fitted model can lead to biased parameter estimates (Table A.3). To conclude, it is crucial to recognize the stratification, especially when the anchor sensitivities varies a lot.

Appendix IV: Supplementary Tables

Table A.1: Estimated odds ratio parameters $\boldsymbol{\beta}$, anchor sensitivity c, and phenotype prevalence q when the phenotype prevalence was equal to $5\%, 10\%, 15\%$, or 20%.

Phenotype prevalence		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	c	q
5%													
	Mean	-2.629	0.208	0.433	0.634	-1.053	-1.472	1.899	-2.112	2.545	2.952	0.500	0.051
	Bias	-0.129	0.008	0.033	0.034	-0.053	-0.072	0.099	-0.112	0.145	0.152	0.0009	-0.0002
	SE	0.561	0.061	0.353	0.122	0.135	0.391	0.240	0.253	0.463	0.357	0.031	0.003
	ESE	0.573	0.062	0.368	0.127	0.144	0.416	0.258	0.269	0.476	0.383	0.031	0.003
10%													
	Mean	1.039	0.207	0.410	0.621	-1.031	-1.444	1.858	-2.067	2.482	2.895	0.500	0.100
	Bias	0.039	0.007	0.010	0.021	-0.031	-0.044	0.058	-0.067	0.082	0.095	0.0004	-0.00008
	SE	0.471	0.046	0.267	0.092	0.101	0.294	0.181	0.189	0.345	0.268	0.021	0.004
	ESE	0.486	0.047	0.271	0.091	0.106	0.301	0.192	0.197	0.350	0.284	0.021	0.004
15%													
	Mean	3.401	0.202	0.406	0.616	-1.023	-1.441	1.842	-2.049	2.461	2.868	0.500	0.148
	Bias	0.101	0.002	0.006	0.016	-0.023	-0.041	0.042	-0.049	0.061	0.068	0.0004	-0.0008
	SE	0.517	0.040	0.232	0.079	0.088	0.255	0.157	0.163	0.299	0.232	0.017	0.005
	ESE	0.520	0.041	0.237	0.082	0.091	0.258	0.166	0.171	0.308	0.237	0.017	0.005
20%													
	Mean	5.511	0.203	0.409	0.609	-1.019	-1.435	1.834	-2.039	2.443	2.856	0.500	0.204
	Bias	0.111	0.003	0.009	0.009	-0.019	-0.035	0.034	-0.039	0.043	0.056	0.0002	-0.0002
	SE	0.583	0.036	0.209	0.072	0.079	0.231	0.143	0.148	0.270	0.211	0.014	0.005
	ESE	0.562	0.036	0.214	0.072	0.078	0.227	0.141	0.144	0.268	0.210	0.014	0.005

Mean: the mean $\hat{\beta}$ estimate; Bias: the difference between the mean $\hat{\beta}$ and the true value of β; SE: the mean asymptotic standard error estimate; ESE: the empirical standard error estimate

Table A.2: Estimated odds ratio parameters $\boldsymbol{\beta}$, anchor sensitivity c, and phenotype prevalence q at 10% phenotype prevalence when $c = 0.2$.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	c	q
Size of training set=10,000												
Mean	1.192	0.220	0.453	0.664	-1.115	-1.573	2.029	-2.243	2.698	3.143	0.201	0.100
Bias	0.192	0.020	0.053	0.064	-0.115	-0.173	0.229	-0.243	0.298	0.343	0.0007	0.0003
SE	0.889	0.085	0.484	0.173	0.202	0.544	0.363	0.383	0.658	0.540	0.017	0.006
ESE	1.422	0.109	0.554	0.189	0.296	0.831	0.737	0.687	1.292	1.085	0.017	0.007
Size of training set=20,000												
Mean	1.064	0.209	0.436	0.628	-1.044	-1.464	1.877	-2.087	2.495	2.925	0.200	0.100
Bias	0.064	0.009	0.036	0.028	-0.044	-0.064	0.077	-0.087	0.095	0.125	0.0005	-0.0003
SE	0.577	0.055	0.320	0.111	0.123	0.354	0.220	0.230	0.416	0.327	0.012	0.004
ESE	0.594	0.056	0.334	0.117	0.125	0.351	0.226	0.237	0.403	0.332	0.012	0.004

Mean: the mean $\hat{\beta}$ over 1000 iterations; Bias: the difference between the mean $\hat{\beta}$ and the true value of β; SE: the mean asymptotic standard error estimate; ESE: the empirical standard error.

Table A.3: Estimated ($\hat{\beta}$, $\hat{c_1}$, $\hat{c_2}$, $\hat{q}$) for stratified conditional independence with $c_1 = 0.2$ and $c_2 = 0.8$ at 10% phenotype prevalence.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	c_1	c_2	q
Varying anchor sensitivity[1]													
Mean	1.046	0.207	0.410	0.622	-1.031	-1.437	1.856	-2.062	2.478	2.882	0.201	0.800	0.100
Bias	0.046	0.007	0.010	0.022	-0.031	-0.037	0.056	-0.062	0.078	0.082	0.001	0.000	-0.000
ASE	0.559	0.041	0.238	0.081	0.089	0.262	0.159	0.165	0.459	0.234	0.025	0.022	0.004
ESE	0.577	0.040	0.243	0.082	0.091	0.258	0.161	0.168	0.476	0.237	0.025	0.022	0.004
Constant anchor sensitivity[2]													
Mean	-2.338	0.148	0.293	0.445	-0.735	-1.025	1.324	-1.471	4.827	2.047	0.712	_	0.080
Bias	-3.338	-0.052	-0.107	-0.155	0.265	0.375	-0.476	0.529	2.427	-0.753	0.212	_	-0.021
ASE	0.363	0.033	0.195	0.064	0.065	0.210	0.116	0.118	0.327	0.168	0.028	_	0.004
ESE	0.504	0.034	0.200	0.071	0.077	0.212	0.138	0.145	0.349	0.209	0.036	_	0.005

Mean: the mean estimates over 1000 iterations; Bias: the difference between the mean estimates and the true values; SE: the mean asymptotic standard error estimate; ESE: the empirical standard error. [1]: the fitted model recognized the variation in anchor sensitivity; [2]: the fitted model failed to recognize the variation in anchor sensitivity.

Table A.4: Variables of interest for primary care patients identifying Penn Medicine hypertension patients with hypokalemia requiring oral supplementation

Variable	$N = 10000$
Mean of systolic blood pressure measured[1]	131.5 (9.5)
Minimum of potassium lab test[1]	3.9 (0.4)
Maximum of potassium lab test[1]	4.7 (0.5)
1st quantile of potassium lab test[1]	4.1 (0.3)
3rd quantile of potassium lab test[1]	27.7 (2.2)
3rd quantile of potassium lab test[1]	4.4 (0.4)
Median of potassium lab test[1]	4.3 (0.3)
Number of low potassium test results[2]	0 (0-44)
Number of Hypokalemia Diagnosis[2]	0 (0-78)
Sum of E87.6 ICD-9 and ICD-10 codes[2]	0 (0-98)
Sum of I10 ICD-9 and ICD-10 codes[2]	19 (0-296)
Sum of I16.1 ICD-9 and ICD-10 codes[2]	2 (0-244)

[1]: Mean (SD), [2]: Median (Minimum-Maximum)

Table A.5: Characteristics of patients in Penn Medicine EHR that had an order for a PA screening laboratory test

	Final population $(N = 6319)$
Age[1]	54 (16)
$Gender$	
Male (%)	2471 (39)
Female (%)	3848 (61)
$Race$	
Caucasian (%)	3270 (52)
African-American (%)	2440 (39)
Other (%)	609 (9)
$Aldosterone$[2] (ng/DL)	8.2 (4.4-14.7)
$Renin$[2] (ng/mL/hr)	0.9 (0.3-2.7)
$Aldo : Renin$[2]	8.0 (2.6-27.8)
$Potassium$[1] (mmol/L)	4.1 (0.6)
$Sodium$[1] (mmol/L)	139 (3.1)
$CO2$[1] (mmol/L)	26.6 (3.4)

[1]: Mean (SD), [2]: Median (IQR)

Table A.6: Dictionary for variables of interest for identifying Penn Medicine hypertension patients with primary aldosteronism

	Variable	Description
Demographics	age	Age when aldosterone or renin test was performed (year)
	gender	Gender
	race	Race (Black/White/Other)
	hisp	Hispanic (Yes/No)
Pre-visit	dbp	Diastolic blood pressure, from office visit closest (<= 14 days) to aldosterone/renin testing
	sbp	Systolic blood pressure, from office visit closest (<= 14 days) to aldosterone/renin testing
	time_bp_to_1st_RAR_yr	Time interval (years) between first office visit with blood pressure recorded to aldosterone/renin test
	time_enc_to_1st_RAR_yr	Time interval (years) between first clinical encounter to aldosterone/renin test
Laboratory data	aldo	Serum aldosterone concentration (ng/dL)
	pra	Plasma renin activity (ng/mL/hr)
	aldo:pra	The aldosterone:renin ratio ((ng Aldosterone/dL)/(ng Angiotensin II/mL/hr))
	test_potassium	Blood potassium concentration (mmol/L)
	test_sodium	Blood sodium concentration (mmol/L)
	test_carbon_dioxide	Blood carbon dioxide concentration (mmol/L)
Encounter	enc_n	Number of clinical encounters
	enc_bp_n	Number of office visits with blood pressure recorded
	time_bp_after_1st_RAR_yr	Time interval (years) between aldosterone/renin test and last office visit with blood pressure
	time_enc_after_1st_RAR_yr	Time interval (years) between aldosterone/renin test and last clinical encounter
Diagnoses	Dx_h2_E26.0_9_n	Sum of the number of encounters with primary aldosteronism diagnosis codes (255.1, 255.10, 255.11, 255.12, E26.0, E26.01, E26.02, E26.09, E26.9)
	Dx_h2_E26.1_8_n	Sum of the number of encounters with other hyperaldosteronism diagnosis codes (255.13, 255.14, E26.1, E26.81, E26.89)
Notes	re_hyperaldo	count of 'hyperaldo' mentions in clinical notes
	re_primaryaldo	count of 'primary aldo' mentioned in the clinical notes
	re_bah	count of 'bah' mentioned in the clinical notes
	re_adrenal_adenoma	count of 'adrenal_adenoma' mentioned in the clinical notes
	re_htn	count of 'hypertension' mentioned in the clinical notes
	re_adrenalectomy	count of 'adrenalectomy' mentioned in the clinical notes

Table A.7: Final phenotyping models for identifying Penn Medicine hypertension patients with primary aldosteronism

	Case set A			Case set B		
	Coefficient	SE	p_value	Coefficient	SE	p_value
Intercept.	-9.06	2.20	0.00	-9.35	2.17	0.00
age	-0.43	0.23	0.06	-0.38	0.20	0.06
gender	0.39	0.37	0.29	0.40	0.33	0.23
race	0.59	0.22	0.01	0.56	0.21	0.01
hisp	-0.12	1.16	0.91	-1.07	1.16	0.35
dbp	-0.31	0.25	0.22	-0.42	0.23	0.07
sbp	0.32	0.27	0.24	0.41	0.24	0.09
time_bp_to_1st_RAR_yr	-0.32	0.31	0.29	-0.16	0.27	0.54
time_enc_to_1st_RAR_yr	-0.20	0.39	0.60	-0.23	0.32	0.48
aldo	0.80	0.39	0.04	0.94	0.38	0.01
pra	0.03	0.61	0.96	-0.05	0.61	0.94
aldo:pra	1.13	0.66	0.09	1.00	0.66	0.13
potassium_indicator	2.89	1.37	0.03	2.96	1.30	0.02
potassium_test	0.43	0.69	0.53	-0.10	0.64	0.88
sodium_indicator	9.08	1.37	0.00	8.85	1.33	0.00
sodium_test	2.82	3.56	0.43	3.78	3.52	0.28
carbon_dioxide_indicator	-10.46	1.37	0.00	-9.82	1.33	0.00
carbon_dioxide_test	-0.33	0.58	0.57	-0.27	0.59	0.64
enc_n	0.80	0.48	0.10	0.25	0.46	0.59
enc_bp_n	-1.53	0.44	0.00	-1.16	0.42	0.01
time_enc_after_1st_RAR_yr	0.53	0.31	0.08	0.28	0.28	0.33
time_bp_after_1st_RAR_yr	-0.12	0.33	0.71	-0.04	0.33	0.90
Dx_h2_E26.0_9_n	0.53	0.16	0.00	1.22	0.23	0.00
Dx_h2_E26.1_8_n	0.17	0.07	0.01	0.17	0.06	0.01
re_hyperaldo	0.38	0.20	0.05	0.39	0.25	0.12
re_primary_aldo	0.13	0.08	0.12	0.54	0.18	0.00
re_bah	1.53	0.53	0.00	2.10	0.49	0.00
re_adrenal_adenoma	0.52	0.16	0.00	0.49	0.14	0.00
re_htn	-0.21	0.32	0.50	-0.24	0.34	0.47
re_adrenalectomy	0.82	0.18	0.00	0.90	0.16	0.00

Table A.8: Estimated anchor sensitivity and prevalence, Mean (95% CI), using stratified vs. non-stratified ML method for identifying PA patients

	Stratified			Non-Stratified	
	c^1	c^2	q	c	q
Set A					
enc_n	0.69 (0.54, 0.84)	0.49 (0.38, 0.6)	0.040 (0.033, 0.048)	0.56 (0.46, 0.65)	0.042 (0.034, 0.050)
enc_bp_n	0.72 (0.52, 0.92)	0.51 (0.4, 0.62)	0.039 (0.031, 0.047)	-	-
time_enc_after_1st_RAR_yr	0.46 (0.31, 0.6)	0.59 (0.49, 0.69)	0.044 (0.035, 0.053)	-	-
time_bp_after_1st_RAR_yr	0.49 (0.33, 0.64)	0.59 (0.47, 0.7)	0.043 (0.034, 0.052)	-	-
Set B					
enc_n	0.76 (0.65, 0.86)	0.52 (0.43, 0.61)	0.049 (0.042, 0.056)	0.62 (0.54, 0.69)	0.051 (0.044, 0.058)
enc_bp_n	0.72 (0.61, 0.84)	0.56 (0.46, 0.65)	0.050 (0.043, 0.057)	-	-
time_enc_after_1st_RAR_yr	0.75 (0.64, 0.87)	0.54 (0.45, 0.63)	0.051 (0.043, 0.058)	-	-
time_bp_after_1st_RAR_yr	0.76 (0.64, 0.87)	0.53 (0.44, 0.62)	0.051 (0.044, 0.058)	-	-

[1]Estimated anchor sensitivity in the 1st strata, i.e., variable value ¡ median

[2]Estimated anchor sensitivity in the 2nd strata, i.e., variable value ¿= median

Table A.9: Estimated regression coefficients β using stratified vs. non-stratified ML method for identifying PA patients

	Set A					Set B				
	Non-S	S_1	S_2	S_3	S_4	Non-S	S_1	S_2	S_3	S_4
(Intercept)	-9.06	-8.59	-8.58	-9.31	-9.11	-9.35	-8.95	-9.17	-9.18	-9.41
age	-0.43	-0.45	-0.44	-0.45	-0.43	-0.38	-0.39	-0.38	-0.36	-0.41
gender	0.39	0.36	0.37	0.39	0.4	0.4	0.4	0.4	0.44	0.42
race	0.59	0.57	0.55	0.6	0.59	0.56	0.56	0.56	0.57	0.58
hisp	-0.12	-0.16	-0.11	-0.12	-0.14	-1.07	-1.31	-1.23	-1.23	-1.27
dbp	-0.31	-0.31	-0.31	-0.31	-0.3	-0.42	-0.42	-0.42	-0.37	-0.39
sbp	0.32	0.3	0.3	0.33	0.32	0.41	0.37	0.41	0.36	0.38
time_bp_to_1st_RAR_yr	-0.32	-0.31	-0.31	-0.32	-0.34	-0.16	-0.13	-0.14	-0.13	-0.1
time_enc_to_1st_RAR_yr	-0.2	-0.19	-0.21	-0.21	-0.19	-0.23	-0.23	-0.23	-0.26	-0.29
aldo	0.8	0.72	0.72	0.85	0.82	0.94	0.81	0.86	0.86	0.88
pra	0.03	0.08	0.04	-0.02	0.01	-0.05	0.04	-0.01	0.02	0.02
aldo:pra	1.13	1.12	1.08	1.08	1.1	1.0	1.01	0.99	1.03	1.05
potassium_indicator	2.89	2.53	2.48	3.15	2.84	2.96	2.44	2.66	2.54	3.08
potassium_test	0.43	0.31	0.29	0.43	0.43	-0.1	-0.22	-0.18	-0.1	-0.12
sodium_indicator	9.08	9.1	9.15	9.1	9.17	8.85	8.99	8.95	8.95	8.76
sodium_test	2.82	2.35	2.34	3.32	2.99	3.78	3.36	3.6	3.43	3.74
carbon_dioxide_indicator	-10.46	-10.43	-10.39	-10.43	-10.36	-9.82	-9.69	-9.72	-9.73	-9.92
carbon_dioxide_test	-0.33	-0.36	-0.34	-0.33	-0.33	-0.27	-0.22	-0.23	-0.25	-0.26
enc_n	0.8	0.95	0.86	0.77	0.77	0.25	0.44	0.3	0.27	0.25
enc_bp_n	-1.53	-1.5	-1.38	-1.55	-1.49	-1.16	-1.15	-1.07	-1.13	-1.2
time_enc_after_1st_RAR_yr	0.53	0.56	0.54	0.35	0.53	0.28	0.31	0.3	0.52	0.26
time_bp_after_1st_RAR_yr	-0.12	-0.19	-0.21	-0.05	-0.22	-0.04	-0.12	-0.09	-0.11	0.22
Dx_h2_E26.0_9_n	0.53	0.53	0.53	0.53	0.52	1.22	1.21	1.23	1.28	1.29
Dx_h2_E26.1_8_n	0.17	0.16	0.16	0.17	0.17	0.17	0.15	0.16	0.16	0.16
re_hyperaldo	0.38	0.39	0.38	0.39	0.38	0.39	0.43	0.42	0.4	0.42
re_primary_aldo	0.13	0.12	0.11	0.14	0.14	0.54	0.54	0.55	0.5	0.49
re_bah	1.53	1.47	1.38	1.49	1.5	2.1	1.93	1.95	2.02	2
re_adrenal_adenoma	0.52	0.53	0.53	0.53	0.51	0.49	0.48	0.49	0.48	0.5
re_htn	-0.21	-0.18	-0.19	-0.22	-0.22	-0.24	-0.18	-0.22	-0.24	-0.24
re_adrenalectomy	0.82	0.79	0.75	0.85	0.83	0.9	0.89	0.88	0.87	0.88

Non-S: non-stratified

S_1-S_4: stratified based on enc_n, enc_bp_n, $time\_enc\_after\_1st\_RAR\_yr$, $time\_bp\_after\_1st\_RAR\_yr$ respectively

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

Appendix I: Influence function of $\widehat{\boldsymbol{\theta}}$

Since the estimates $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}^{\mathrm{T}}, c)^{\mathrm{T}}$ are classical MLE, we know the large sample property of $\boldsymbol{\theta}$ is $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to \mathrm{Normal}(\mathbf{0}, \mathbf{V})$ in distribution when $N \to \infty$, where

$$\mathbf{V} = \left\{ -E\frac{1}{N}\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} \right\}^{-1}.$$

An obvious estimator of $\mathbf{V}$ is

$$\widehat{\mathbf{V}} = \left\{ -\frac{1}{N}\frac{\partial^2 l(\widehat{\boldsymbol{\theta}})}{\partial\widehat{\boldsymbol{\theta}}\partial\widehat{\boldsymbol{\theta}}^{\mathrm{T}}} \right\}^{-1}.$$

Let $\boldsymbol{\phi}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c) = \{\boldsymbol{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c)^{\mathrm{T}}, \phi_c(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c)\}^{\mathrm{T}}$ be the score function, where

$$\begin{aligned}
\boldsymbol{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c) &= S_i\frac{\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})}{P(\mathbf{X}_i;\boldsymbol{\beta})} - (1-S_i)\frac{c\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})}{1-cP(\mathbf{X}_i;\boldsymbol{\beta})} \\
\phi_c(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c) &= S_i\frac{1}{c} - (1-S_i)\frac{P(\mathbf{X}_i;\boldsymbol{\beta})}{1-cP(\mathbf{X}_i;\boldsymbol{\beta})}.
\end{aligned}$$

Let

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11}, & \mathbf{A}_{12} \\ \mathbf{A}_{12}^{\mathrm{T}}, & A_{22} \end{pmatrix},$$

where

$$\begin{aligned}
\mathbf{A}_{11} &= E\left\{ S_i\frac{\mathbf{P}''_{\boldsymbol{\beta}\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})}{P(\mathbf{X}_i;\boldsymbol{\beta})} - S_i\frac{\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})^{\mathrm{T}}}{P(\mathbf{X}_i;\boldsymbol{\beta})^2} \right. \\
&\quad \left. -(1-S_i)\frac{c\mathbf{P}''_{\boldsymbol{\beta}\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})}{1-cP(\mathbf{X}_i;\boldsymbol{\beta})} + (1-S_i)\frac{c^2\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})^{\mathrm{T}}}{\{1-cP(\mathbf{X}_i;\boldsymbol{\beta})\}^2} \right\}, \\
\mathbf{A}_{12} &= E\left\{ -(1-S_i)\frac{\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})}{1-cP(\mathbf{X}_i;\boldsymbol{\beta})} - (1-S_i)\frac{cP(\mathbf{X}_i;\boldsymbol{\beta})\mathbf{P}'_{\boldsymbol{\beta}}(\mathbf{X}_i;\boldsymbol{\beta})}{\{1-cP(\mathbf{X}_i;\boldsymbol{\beta})\}^2} \right\}, \\
A_{22} &= E\left\{ S_i\frac{-1}{c^2} - (1-S_i)\frac{P(\mathbf{X}_i;\boldsymbol{\beta})^2}{\{1-cP(\mathbf{X}_i;\boldsymbol{\beta})\}^2} \right\}.
\end{aligned}$$

Then we have the expansion

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = N^{-1} \sum_{i=1}^{N} \boldsymbol{\psi}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) + o_p(N^{-1/2}), \tag{B.1}$$

where $\boldsymbol{\psi}(\mathbf{X}, S, \boldsymbol{\theta})$ is the influence function of $\boldsymbol{\theta}$, with $\boldsymbol{\psi}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) = -\mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c)$. For later use, we extract the $\boldsymbol{\beta}$ part from above and write

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = N^{-1} \sum_{i=1}^{N} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c) + o_p(N^{-1/2})$$

and extract the c part and write

$$\widehat{c} - c = N^{-1} \sum_{i=1}^{N} \psi_c(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c) + o_p(N^{-1/2}).$$

Appendix II: Property of T^q when the fitted model is correct.

When the phenotype prevalence q is known, the statistic T^q is constructed as

$$T^q = \mathbf{U}_K^{\mathrm{T}} \mathbf{V}_K^{-1} \mathbf{U}_K, \tag{B.2}$$

where $\mathbf{U}_K$ is a length K vector whose lth element ($l = 1, ..., K-1$) is

$$U_{K,l} = \widehat{p}^n\{Y = 1 | P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \in I_l, S = 0\} - \widehat{p}^m\{Y = 1 | P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) \in I_l, S = 0\},$$

and Kth element is $U_{K,K} = \widehat{q} - q = N^{-1} \sum_{i=1}^{N} P(\mathbf{X}; \widehat{\boldsymbol{\beta}}) - q$.

$\mathbf{V}_K$ is the variance-covariance matrix of $\mathbf{U}_K$, with element

$$
\begin{aligned}
V_{K,l,j} &= cov(U_{K,l}, U_{K,j}) \\
&= \mathsf{cov}[\hat{p}^n\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l, S = 0\} - \hat{p}^m\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l; S = 0\}, \\
&\quad \hat{p}^n\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_j, S = 0\} - \hat{p}^m\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_j; S = 0\}] \\
V_{K,l,K} &= cov(U_{K,l}, U_{K,K}) \\
&= \mathsf{cov}[\hat{p}^n\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l, S = 0\} - \hat{p}^m\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l; S = 0\}, \widehat{q} - q] \\
V_{K,K,K} &= cov(U_{K,K}, U_{K,K}) \\
&= \mathsf{cov}(\widehat{q} - q, \widehat{q} - q)
\end{aligned}
$$

for $l = 1, ..., K-1, j = 1, ..., K-1$.

Since

$$
\begin{aligned}
& \hat{p}^n\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l, S = 0\} - \hat{p}^m\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l; S = 0\} \\
= \; & \hat{p}^n\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l, S = 0\} - p\{Y = 1 | P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\} \\
- \; & [\hat{p}^m\{Y = 1 | P(\mathbf{X}; \hat{\boldsymbol{\beta}}) \in I_l, S = 0\} - p\{Y = 1 | P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}] \\
= \; & U^n - U^m,
\end{aligned}
$$

where

$$
\begin{aligned}
U^n &= \hat{p}^n\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&= \hat{p}^n\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0\} - \hat{p}^n\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&+ \hat{p}^n\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&= R_1^n + R_2^n \\
U^m &= \hat{p}^m\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&= \hat{p}^m\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0\} - \hat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&+ \hat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&= R_1^m + R_2^m,
\end{aligned}
$$

with

$$
\begin{aligned}
R_1^n &= \hat{p}^n\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0\} - \hat{p}^n\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
R_2^n &= \hat{p}^n\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
R_1^m &= \hat{p}^m\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0\} - \hat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
R_2^m &= \hat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\}.
\end{aligned}
$$

Thus know that the variance of $\mathbf{U}_K$ is composed of two parts, one from parameter estimation, e.g., R_1^n and R_1^m, the other one from the construction of p^n and p^m, e.g., R_2^n and R_2^m. Below we show the detailed estimation of $\mathbf{V}_K$ when the risk model was developed from positive-only data for illustration. It can be estimated similarly even if the model was developed externally of the positive-only data.

For notational brevity, let $\widehat{h} \equiv N^{-1}\sum_{i=1}^N I(S_i=1)$, $p_{l0} \equiv p\{P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\}$, $\widehat{p}_{l0} \equiv N^{-1}\sum_{i=1}^N I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in$

$I_l, S_i = 0\}$, $p_{l1} \equiv p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 1\}$, $\widehat{p}_{l1} \equiv N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\boldsymbol{\beta}) \in I_l, S_i = 1\}$. Then

$$\begin{aligned}
R_1^n &= \widehat{p}^n\{Y = 1|P(\mathbf{X};\hat{\boldsymbol{\beta}}) \in I_l, S = 0, q\} - \widehat{p}^n\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0, q\} \\
&= \frac{\partial \widehat{p}^n\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}^*) \in I_l, S = 0, q^*\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \left(E\left[\frac{\partial p^n\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0, q\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right] + o_p(1)\right)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= N^{-1}\sum_{i=1}^{N} \mathbf{a}_{ln}^{\mathrm{T}} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c) + o_p(N^{-1}),
\end{aligned}$$

where

$$\mathbf{a}_{ln} \equiv E\left[\frac{\partial p\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0, q\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right],$$

and $\boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}, S, \boldsymbol{\beta}, c)$ is the influence function of $\boldsymbol{\beta}$ as shown in equation (B.2) of Appendix A. In addition,

$$\begin{aligned}
R_2^n &= \widehat{p}^n\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\} - p\{Y = 1|P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\} \\
&= \frac{\{q - N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\boldsymbol{\beta}) \in I_l, S_i = 1\}]}{\{N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\boldsymbol{\beta}) \in I_l, S_i = 0\}]} \\
&\quad - \frac{(q - h)p\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 1\}}{hp\{P(\mathbf{X};\boldsymbol{\beta}) \in I_l, S = 0\}} \\
&= \frac{(q - \widehat{h})\widehat{p}_{l1}}{\widehat{h}\widehat{p}_{l0}} - \frac{(q - h)p_{l1}}{hp_{l0}} \\
&= \frac{q - h}{hp_{l0}}(\widehat{p}_{l1} - p_{l1}) - \frac{(q - h)p_{l1}}{hp_{l0}^2}(\widehat{p}_{l0} - p_{l0}) - \frac{qp_{l1}}{h^2 p_{l0}}(\widehat{h} - h) + O_p(N^{-1}) \\
&= N^{-1}\sum_{i=1}^{N} \psi_{ln}(\mathbf{X}_i, S_i, \boldsymbol{\beta}) + o_p(N^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
\psi_{ln}(\mathbf{X}_i, S_i, \boldsymbol{\beta}) \equiv\ & \frac{q - h}{hp_{l0}}[I\{P(\mathbf{X}_i;\boldsymbol{\beta}) \in I_l, S_i = 1\} - p_{l1}] \\
& - \frac{(q - h)p_{l1}}{hp_{l0}^2}[I\{P(\mathbf{X}_i;\boldsymbol{\beta}) \in I_l, S_i = 0\} - p_{l0}] \\
& - \frac{qp_{l1}}{h^2 p_{l0}}\{I(S_i = 1) - h\}.
\end{aligned} \tag{B.3}$$

Thus,

$$\begin{aligned} U^n &= \widehat{p}^n\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0, \hat{q}\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, q\} \\ &= N^{-1}\sum_{i=1}^{N}\phi_{ln}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) + o_p(N^{-1}), \end{aligned}$$

where

$$\phi_{ln}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) = \psi_{ln}(\mathbf{X}_i, S_i, \boldsymbol{\beta}) + \mathbf{a}_{ln}^{\mathrm{T}}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c), \tag{B.4}$$

which has mean zero.

On the other hand, applying Taylor's Theorem, we have

$$\begin{aligned} R_1^m &= \widehat{p}^m\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0, \widehat{c}\} - \widehat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, c\} \\ &= \frac{\partial\widehat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta}^*)\in I_l, S=0, c^*\}}{\partial\boldsymbol{\theta}^{\mathrm{T}}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \\ &= \left(\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\frac{\sum_{i=1}^{N}(1-c^*)P(\mathbf{X}_i,\boldsymbol{\beta}^*)I\{P(\mathbf{X}_i;\boldsymbol{\beta}^*)\in I_l\}}{\sum_{i=1}^{N}\{1-c^*P(\mathbf{X}_i,\boldsymbol{\beta}^*)\}I\{P(\mathbf{X}_i;\boldsymbol{\beta}^*)\in I_l\}} + o_p(1)\right)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \\ &= \left(E\left[\frac{\partial p^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, c\}}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\right] + o_p(1)\right)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \\ &= N^{-1}\sum_{i=1}^{N}\mathbf{a}_{lm}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) + o_p(N^{-1}), \end{aligned}$$

where

$$\mathbf{a}_{lm} \equiv E\left[\frac{\partial p^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, c\}}{\partial\boldsymbol{\theta}}\right], \tag{B.5}$$

and $\boldsymbol{\psi}(\mathbf{X}, S, \boldsymbol{\theta})$ is the influence function of $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, c)$ as shown in equation (B.1) of Appendix A.

Further,

$$
\begin{aligned}
R_2^m &= \widehat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, c\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&= \frac{N^{-1}\sum_{i=1}^N (1-c)P(\mathbf{X}_i,\boldsymbol{\beta})I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\}}{N^{-1}\sum_{i=1}^N\{1-cP(\mathbf{X}_i,\boldsymbol{\beta})\}I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\}} - \frac{\int_{I_l}(1-c)P(\mathbf{X},\boldsymbol{\beta})dF(\mathbf{X})}{\int_{I_l}\{1-cP(\mathbf{X},\boldsymbol{\beta})\}dF(\mathbf{X})} \\
&= \frac{[N^{-1}\sum_{i=1}^N (1-c)P(\mathbf{X}_i,\boldsymbol{\beta})I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\} - \int_{I_l}(1-c)P(\mathbf{X},\boldsymbol{\beta})dF(\mathbf{X})]}{\int_{I_l}\{1-cP(\mathbf{X},\boldsymbol{\beta})\}dF(\mathbf{X})} \\
&\quad -\frac{p^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, c\}}{\int_{I_l}\{1-cP(\mathbf{X},\boldsymbol{\beta})\}dF(\mathbf{X})} \\
&\quad \times[N^{-1}\sum_{i=1}^N\{1-cP(\mathbf{X}_i,\boldsymbol{\beta})\}I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\} - \int_{I_l}\{1-cP(\mathbf{X},\boldsymbol{\beta})\}dF(\mathbf{X})] + o_p(N^{-1}) \\
&= N^{-1}\sum_{i=1}^N [(1-c)P(\mathbf{X}_i,\boldsymbol{\beta})I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\} - p^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, c\} \\
&\quad \times\{1-cP(\mathbf{X}_i,\boldsymbol{\beta})\}I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\}]\,/[\int_{I_l}\{1-cP(\mathbf{X},\boldsymbol{\beta})\}dF(\mathbf{X})] + o_p(N^{-1}) \\
&= N^{-1}\sum_{i=1}^N \psi_{lm}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c),
\end{aligned}
$$

where

$$
\begin{aligned}
\psi_{lm}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c) \equiv\ & [(1-c)P(\mathbf{X}_i,\boldsymbol{\beta})I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\} - p\{Y=1|P(\mathbf{X},\boldsymbol{\beta})\in I_l, S=0, c\} \\
& \times\{1-cP(\mathbf{X}_i,\boldsymbol{\beta})\}I\{P(\mathbf{X}_i;\boldsymbol{\beta})\in I_l\}]\,/[\int_{I_l}\{1-cP(\mathbf{x},\boldsymbol{\beta})\}dF(\mathbf{X})].
\end{aligned} \tag{B.6}
$$

Then

$$
\widehat{p}^m\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0, c\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} = N^{-1}\sum_{i=1}^N \psi_{lm}(\mathbf{X}_i, S_i, \boldsymbol{\beta}, c).
$$

Thus,

$$
\begin{aligned}
U^m &= \widehat{p}^m\{Y=1|P(\mathbf{X};\hat{\boldsymbol{\beta}})\in I_l, S=0, \widehat{c}\} - p\{Y=1|P(\mathbf{X};\boldsymbol{\beta})\in I_l, S=0\} \\
&= N^{-1}\sum_{i=1}^N \phi_{lm}(\mathbf{X}_i, x_i, \boldsymbol{\theta}) + o_p(N^{-1}),
\end{aligned}
$$

where

$$
\phi_{lm}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) \equiv \psi_{lm}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) + \mathbf{a}_{lm}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{X}_i, S_i, \boldsymbol{\theta}), \tag{B.7}
$$

which has mean zero.

Further, by Taylor expansion, we have

$$\begin{aligned}
& \widehat{q} - q \\
= \;& N^{-1} \sum_{i=1}^{N} P(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}) - q \\
= \;& N^{-1} \sum_{i=1}^{N} \{P(\mathbf{X}_i; \boldsymbol{\beta}) - q\} + N^{-1} \sum_{i=1}^{N} P'_{\boldsymbol{\beta}}(\mathbf{X}_i; \boldsymbol{\beta})^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(N^{-1/2}) \\
= \;& N^{-1} \sum_{i=1}^{N} \{P(\mathbf{X}_i; \boldsymbol{\beta}) - q\} + N^{-1} \sum_{i=1}^{N} E\{P'_{\boldsymbol{\beta}}(\mathbf{X}_i; \boldsymbol{\beta})\} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i, \boldsymbol{\theta}) + o_p(N^{-1/2}) \\
= \;& N^{-1} \sum_{i=1}^{N} [P(\mathbf{X}_i; \boldsymbol{\beta}) - q + E\{P'_{\boldsymbol{\beta}}(\mathbf{X}_i; \boldsymbol{\beta})\} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i, \boldsymbol{\theta})] + o_p(N^{-1/2}) \\
= \;& N^{-1} \sum_{i=1}^{N} \phi_K(\mathbf{X}, S, \boldsymbol{\theta}) + o_p(N^{-1/2}),
\end{aligned}$$

where

$$\phi_K(\mathbf{X}, S, \boldsymbol{\theta}) \equiv [P(\mathbf{X}; \boldsymbol{\beta}) - q + E\{\frac{\partial P(\mathbf{X}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\}^{\mathrm{T}} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}, S, \boldsymbol{\theta})], \tag{B.8}$$

which has mean zero.

Thus, we have

$$\begin{aligned}
U_{K,l} &\equiv N^{-1} \sum_{i=1}^{N} \{\phi_{ln}(\mathbf{X}, S, \boldsymbol{\theta}) - \phi_{lm}(\mathbf{X}, S, \boldsymbol{\theta})\} + o_p(N^{-1}), \text{ for } l = 1, \ldots, K-1, \\
U_{K,K} &\equiv N^{-1} \sum_{i=1}^{N} \phi_K(\mathbf{X}, S, \boldsymbol{\theta}) + o_p(N^{-1/2}),
\end{aligned} \tag{B.9}$$

with ϕ_{ln}, ϕ_{lm} and ϕ_K as expressed in (B.4), (B.7), (B.8) respectively, which has mean zero. Thus the asymptotic variance of $\mathbf{U}_K$, $\mathbf{V}_K$ has element as

$$\begin{aligned}
V_{K,l,j} &= E[\{\phi_{ln}(\mathbf{X}, S, \boldsymbol{\theta}) - \phi_{lm}(\mathbf{X}, S, \boldsymbol{\theta})\}\{\phi_{jn}(\mathbf{X}, S, \boldsymbol{\theta}) - \phi_{jm}(\mathbf{X}, S, \boldsymbol{\theta})\}], \\
V_{K,l,K} &= E[\{\phi_{ln}(\mathbf{X}, S, \boldsymbol{\theta}) - \phi_{lm}(\mathbf{X}, S, \boldsymbol{\theta})\}\phi_K(\mathbf{X}, S, \boldsymbol{\theta})], \\
V_{K,K,K} &= E[\phi_K(\mathbf{X}, S, \boldsymbol{\theta})\phi_K(\mathbf{X}, S, \boldsymbol{\theta})],
\end{aligned} \tag{B.10}$$

for $l = 1, ..., K-1, j = 1, ..., K-1$. Thus the test statistic T^q asymptotically follows a χ^2 distribution with K degrees of freedom under H_0, i.e. when the fitted model is correct. We can form $\widehat{\mathbf{V}}_K$ by using the sample average to replace E in $V_{K,l,j}, V_{K,l,K}, V_{K,K,K}$ and plug in the estimated parameter values. When the model was developed externally of the positive-only data, $\mathbf{V}_K$ can be estimated similarly implementing the appropriate influence function, only that ψ_{lm} as shown in equation (B.7) should be adjusted according to the fitted model.

Appendix III: Property of T^{nq} when the fitted model is correct.

When q is unknown,

$$T^{nq} = \mathbf{U}_{K-1}^{\mathrm{T}} \mathbf{V}_{K-1}^{-1} \mathbf{U}_{K-1},$$

where $\mathbf{U}_{K-1}$ is a length $K-1$ vector whose lth element is

$$U_{K-1,l} = \widehat{p}^n\{Y = 1 | P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\} - \widehat{p}^m\{Y = 1 | P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\},$$

and $\mathbf{V}_{K-1}$ is the asymptotic variance of $\mathbf{U}_{K-1}$. Here, $\mathbf{U}_{K-1}$ has the same form as (B.9), only that $\phi_{ln}(\mathbf{X}, S, \boldsymbol{\theta})$ now includes an additional term involving $\{P(\mathbf{X}; \boldsymbol{\beta}) - q\}$ because of the estimation of $\hat{q}$ and becomes,

$$\begin{aligned}
\phi^*_{ln}(\mathbf{X}, S, \boldsymbol{\theta}) \equiv\ & E\left[\frac{\partial p\{Y = 1 | P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right]^{\mathrm{T}} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}, S, \boldsymbol{\theta}) \\
& + \frac{q - p(S=1)}{hp\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}}[I\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 1\} - p\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 1\}] \\
& - \frac{\{q - p(S=1)\}p\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 1\}}{p(S=1)p^2\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}}[I\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\} - p\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}] \\
& - \frac{qp\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 1\}}{p^2(S=1)p\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}}\{I(S=1) - p(S=1)\} \\
& + \frac{p\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 1\}}{p(S=1)p\{P(\mathbf{X}; \boldsymbol{\beta}) \in I_l, S = 0\}}\{P(\mathbf{X}; \boldsymbol{\beta}) - q\}.
\end{aligned}$$

On the other hand, $\phi_{lm}(\mathbf{X}, S, \boldsymbol{\theta})$ remains the same as in (B.7). The (l, j)th element of the $\mathbf{V}_{K-1}$ remains the same as shown in (B.10) with $\phi_{ln}(\mathbf{X}, S, \boldsymbol{\theta})$ and $\phi_{jn}(\mathbf{X}, S, \boldsymbol{\theta})$ replaced by $\phi^*_{ln}(\mathbf{X}, S, \boldsymbol{\theta})$ and $\phi^*_{jn}(\mathbf{X}, S, \boldsymbol{\theta})$ respectively. Thus the test statistic T^{nq} asymptotically follows a χ^2 distribution with $K-1$ degrees of freedom when the fitted model is correct. Again we can form $\widehat{\mathbf{V}}_{K-1}$ by using the sample average to replace E in (B.10) and plug in the estimated parameter values.

Appendix IV: Statistical measures for quantifying predictive accuracy

By definition anchor-labeled patients are definitive cases, thus we assess model performance only among the unlabeled patients. Taking advantage of the unique properties of the anchor variable, the accuracy measures can be estimated nonparametrically using positive-only data.

For notational brevity, let $p_{v0} \equiv p\{P(\mathbf{X}_i;\boldsymbol{\beta}) > v, S_i = 0\}$, $\widehat{p}_{v0} \equiv N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\boldsymbol{\beta}) > v, S_i = 0\}$, $p_{v1} \equiv p\{P(\mathbf{X}_i;\boldsymbol{\beta}) > v, S_i = 1\}$, $\widehat{p}_{v1} \equiv N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\boldsymbol{\beta}) > v, S_i = 1\}$, $p^*_{v0} \equiv p\{P(\mathbf{X}_i;\boldsymbol{\beta}) < v, S_i = 0\}$, $\widehat{p}^*_{v0} \equiv N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\boldsymbol{\beta}) < v, S_i = 0\}$, $p^*_{v1} \equiv p\{P(\mathbf{X}_i;\boldsymbol{\beta}) < v, S_i = 1\}$, $\widehat{p}^*_{v1} \equiv N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\boldsymbol{\beta}) < v, S_i = 1\}$.

The true positive rate at decision threshold v (TPR_v) is defined as

$$\begin{aligned} TPR_v &= p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 1, S = 0\} \\ &= p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 1, S = 1\} \\ &= p\{P(\mathbf{X};\boldsymbol{\beta}) > v | S = 1\} \\ &= \frac{p_{v1}}{h}, \end{aligned}$$

thus TPR_v for the unlabeled can be estimated using anchor-positive cases as

$$\widehat{TPR}_v = \frac{N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S_i = 1\}}{N^{-1}\sum_{i=1}^{N} I(S_i = 1)}.$$

The false positive rate FPR_v is defined as

$$\begin{aligned} FPR_v &= p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 0, S = 0\} \\ &= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\} - p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0, Y = 1\}}{p(Y = 0, S = 0)} \\ &= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\} - p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 1, S = 0\}p\{Y = 1, S = 0\}}{p(Y = 0)} \\ &= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\} - p\{P(\mathbf{X};\boldsymbol{\beta}) > v | S = 1\}[p\{Y = 1\} - p\{Y = 1, S = 1\}]}{p(Y = 0)} \\ &= \frac{p_{v0} - p_{v1}(q - h)/h}{1 - q} \end{aligned}$$

Therefore, it can be estimated similarly as for TPR_v

$$\begin{aligned}\widehat{FPR_v} &= \frac{N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S_i = 0\}}{1 - N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})} \\ &- \frac{[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S_i = 1\}]\{N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) - N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}}{\{N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}\{1 - N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}})\}}.\end{aligned}$$

Similarly, the positive predictive value PPV_v is defined as

$$\begin{aligned}PPV_v &= p\{Y = 1 | P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\} \\ &= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) > v | Y = 1, S = 0\} p\{Y = 1, S = 0\}}{p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\}} \\ &= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) > v | S = 1\}[p\{Y = 1\} - p\{Y = 1, S = 1\}]}{p\{P(\mathbf{X};\boldsymbol{\beta}) > v, S = 0\}} \\ &= \frac{p_{v1}(q - h)}{h p_{v0}},\end{aligned}$$

and further estimated as

$$\widehat{PPV_v} = \frac{[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S_i = 1\}]\{N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) - N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}}{\{N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) > v, S_i = 0\}]}.$$

The negative predictive value NPV_v is defined as

$$\begin{aligned}NPV_v &= p\{Y = 0 | P(\mathbf{X};\boldsymbol{\beta}) < v, S = 0\} \\ &= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) < v, S = 0\} - p\{P(\mathbf{X};\boldsymbol{\beta}) < v, S = 0, Y = 1\}}{p\{P(\mathbf{X};\boldsymbol{\beta}) < v, S = 0\}} \\ &= \frac{p\{P(\mathbf{X};\boldsymbol{\beta}) < v, S = 0\} - p\{P(\mathbf{X};\boldsymbol{\beta}) < v | S = 1\} p\{Y = 1, S = 0\}}{p\{P(\mathbf{X};\boldsymbol{\beta}) < v, S = 0\}} \\ &= 1 - \frac{p^*_{v1}(q - h)}{h p^*_{v0}},\end{aligned}$$

and furter estimated as

$$\widehat{NPV_v} = 1 - \frac{[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) < v, S_i = 1\}]\{N^{-1}\sum_{i=1}^{N} P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) - N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}}{\{N^{-1}\sum_{i=1}^{N} I(S_i = 1)\}[N^{-1}\sum_{i=1}^{N} I\{P(\mathbf{X}_i;\widehat{\boldsymbol{\beta}}) < v, S_i = 0\}]}.$$

Lastly, AUC can be estimated as

$$\widehat{AUC} = \int \widehat{TPR}_v d\widehat{FPR}_v.$$

The asymptotic properties of the estimated accuracy measures are provided below:

$$
\begin{aligned}
& \widehat{TPR}_v - TPR_v \\
= \; & \widehat{TPR}_v(\widehat{\boldsymbol{\beta}}) - \widehat{TPR}_v(\boldsymbol{\beta}) + \widehat{TPR}_v(\boldsymbol{\beta}) - TPR_v \\
= \; & \widehat{TPR}_v(\widehat{\boldsymbol{\beta}}) - \widehat{TPR}_v(\boldsymbol{\beta}) + \frac{\widehat{p}_{v1}}{\hat{h}} - \frac{p_{v1}}{h} \\
= \; & \frac{\partial p\{P(\mathbf{X};\boldsymbol{\beta}) > v \mid S = 1\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{h}(\widehat{p}_{v1} - p_{v1}) - \frac{p_{v1}}{h^2}(\hat{h} - h) + o_p(N^{-1}) \\
= \; & N^{-1}\sum_{i=1}^{N}\{\frac{1}{h}[I\{P(\mathbf{X}_i;\boldsymbol{\beta}) > v, S_i = 1\} - p_{v1}] - \frac{p_{v1}}{h^2}\{I(S_i = 1) - h\} \\
& + \frac{\partial p\{P(\mathbf{X};\boldsymbol{\beta}) > v \mid S = 1\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i)\} + o_p(N^{-1}).
\end{aligned}
$$

$$
\begin{aligned}
& \widehat{FPR}_v - FPR_v \\
= \; & \frac{\partial p\{P(\mathbf{X};\boldsymbol{\beta}) > v \mid Y = 0, S = 0\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
& + \frac{1}{1-q}(\widehat{p}_{v0} - p_{v0}) - \frac{q-h}{h(1-q)}(\widehat{p}_{v1} - p_{v1}) + \frac{p_{v1}q}{h^2(1-q)}(\hat{h} - h) + \frac{h(p_{v0} + p_{v1}) - p_{v1}}{h(1-q)^2}(\hat{q} - q) + o_p(N^{-1}) \\
= \; & N^{-1}\sum_{i=1}^{N}\{\frac{1}{1-q}[I\{P(\mathbf{X}_i;\boldsymbol{\beta}) > v, S_i = 0\} - p_{v0}] - \frac{q-h}{h(1-q)}[I\{P(\mathbf{X}_i;\boldsymbol{\beta}) > v, S_i = 1\} - p_{v1}] \\
& + \frac{p_{v1}q}{h^2(1-q)}\{I(S_i = 1) - h\} + \frac{h(p_{v0} + p_{v1}) - p_{v1}}{h(1-q)^2}\{P(\mathbf{X}_i;\boldsymbol{\beta}) - q\} \\
& + \frac{\partial p\{P(\mathbf{X};\boldsymbol{\beta}) > v \mid Y = 0, S = 0\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i)\} + o_p(N^{-1}).
\end{aligned}
$$

$$
\begin{aligned}
& \widehat{PPV}_v - PPV_v \\
= \; & \frac{\partial p\{Y=1 \mid P(\mathbf{X};\boldsymbol{\beta})>v, S=0\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
& + \frac{q-h}{hp_{v0}}(\widehat{p}_{v1}-p_{v1}) - \frac{p_{v1}(q-h)}{hp_{v0}^2}(\widehat{p}_{v0}-p_{v0}) - \frac{p_{v1}q}{h^2 p_{v0}}(\hat{h}-h) + \frac{p_{v1}}{hp_{v0}}(\hat{q}-q) + o_p(N^{-1}) \\
= \; & N^{-1}\sum_{i=1}^{N}\{\frac{q-h}{hp_{v0}}[I\{P(\mathbf{X}_i;\boldsymbol{\beta})>v, S_i=1\}-p_{v1}] - \frac{p_{v1}(q-h)}{hp_{v0}^2}[I\{P(\mathbf{X}_i;\boldsymbol{\beta})>v, S_i=0\}-p_{v0}] \\
& - \frac{p_{v1}q}{h^2 p_{v0}}\{I(S_i=1)-h\} + \frac{p_{v1}}{hp_{v0}}\{P(\mathbf{X}_i;\boldsymbol{\beta})-q\} \\
& + \frac{\partial p\{Y=1 \mid P(\mathbf{X};\boldsymbol{\beta})>v, S=0\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i)\} + o_p(N^{-1}).
\end{aligned}
$$

$$
\begin{aligned}
& \widehat{NPV}_v - NPV_v \\
= \; & \frac{\partial p\{Y=0 \mid P(\mathbf{X};\boldsymbol{\beta})<v, S=0\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
& - \frac{q-h}{hp_{v0}^*}(\widehat{p}_{v1}^*-p_{v1}^*) + \frac{p_{v1}^*(q-h)}{hp_{v0}^{*2}}(\widehat{p}_{v0}^*-p_{v0}^*) + \frac{p_{v1}^*q}{h^2 p_{v0}}(\hat{h}-h) - \frac{p_{v1}^*}{hp_{v0}^*}(\hat{q}-q) + o_p(N^{-1}) \\
= \; & N^{-1}\sum_{i=1}^{N}\{-\frac{q-h}{hp_{v0}^*}[I\{P(\mathbf{X}_i;\boldsymbol{\beta})<v, S_i=1\}-p_{v1}^*] + \frac{p_{v1}^*(q-h)}{hp_{v0}^{*2}}[I\{P(\mathbf{X}_i;\boldsymbol{\beta})<v, S_i=0\}-p_{v0}^*] \\
& + \frac{p_{v1}^*q}{h^2 p_{v0}^*}\{I(S_i=1)-h\} - \frac{p_{v1}^*}{hp_{v0}^*}\{P(\mathbf{X}_i;\boldsymbol{\beta})-q\} \\
& + \frac{\partial p\{Y=0 \mid P(\mathbf{X};\boldsymbol{\beta})<v, S=0\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\mathbf{X}_i, S_i)\} + o_p(N^{-1}).
\end{aligned}
$$

Lastly,

$$
\begin{aligned}
& \widehat{AUC} - AUC \\
= \; & \int_{v=0}^{1}(\widehat{TPR}_v - TPR_v)d(\widehat{FPR}_v - FPR_v) \\
& + \int_{v=0}^{1}(\widehat{TPR}_v - TPR_v)dFPR_v + \int_{v=0}^{1} TPR_v d(\widehat{FPR}_v - FPR_v) \\
= \; & N^{-1}\sum_{i=1}^{N}\left\{\int_{v=0}^{1}\psi_{TPR}(\mathbf{X}_i, S_i)\frac{\partial FPR_v}{\partial v}dv + \int_{v=0}^{1} TPR_v \frac{\partial \psi_{FPR}(\mathbf{X}_i, S_i)}{\partial v}dv\right\} + o_p(N^{-1}).
\end{aligned}
$$

Appendix VI: Supplementary tables

Table B.1: Calibration slope when N=50000. Benchmark: calibration slope estimated using fully observed Y; Positive-only: calibration slope estimated using positive-only data; Correct: fitted model is correctly specified; Minor: fitted model missing weak predictors (X_1, X_2, X_3); Moderate: fitted model missing moderate predictors (X_4, X_5, X_6); Severe: fitted model missing weak and moderate predictors $(X_1 - X_6)$.

	Benchmark				Positive-only			
	Correct	Minor	Moderate	Severe	Correct	Minor	Moderate	Severe
Slope b	1.000	0.817	0.363	0.351	1.001	0.827	0.378	0.381

Table B.2: Asymptotic standard errors $\times 10$ vs. empirical standard errors $\times 10$ for estimated predictive accuracy measures among unlabeled patients using positive-only data for $N = 10,000$ and $N = 80,000$. Results based on mean over 1000 replications.

		Asymptotic Standard Error					Empirical Standard Error				
	Cutoff	$\widehat{\text{PPV}}$	$\widehat{\text{NPV}}$	$\widehat{\text{TPR}}$	$\widehat{\text{FPR}}$	$\widehat{\text{AUC}}$	$\widehat{\text{PPV}}$	$\widehat{\text{NPV}}$	$\widehat{\text{TPR}}$	$\widehat{\text{FPR}}$	$\widehat{\text{AUC}}$
$N = 10,000$	0.2	0.228	0.009	0.068	0.047	0.012	0.219	0.006	0.049	0.044	0.017
	0.4	0.213	0.014	0.113	0.032	-	0.170	0.010	0.087	0.024	-
	0.6	0.213	0.019	0.163	0.027	-	0.133	0.015	0.130	0.015	-
	0.8	0.214	0.024	0.215	0.022	-	0.110	0.021	0.193	0.011	-
$N = 80,000$	0.2	0.076	0.002	0.018	0.015	0.004	0.076	0.002	0.017	0.015	0.006
	0.4	0.065	0.004	0.033	0.009	-	0.060	0.003	0.029	0.008	-
	0.6	0.057	0.006	0.050	0.006	-	0.047	0.005	0.046	0.005	-
	0.8	0.049	0.007	0.070	0.005	-	0.040	0.007	0.066	0.004	-

APPENDIX C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Appendix I: Supplementary Tables

Table C.1: Dictionary for predictors included in the final risk prediction models for primary aldosteronism

	Variable	Description
Demographics	gender	Gender
	race	Race (Black/White/Other)
	hisp	Hispanic (Yes/No)
Pre-visit	dbp	Diastolic blood pressure, from office visit closest (<= 14 days) to aldosterone/renin testing
	sbp	Systolic blood pressure, from office visit closest (<= 14 days) to aldosterone/renin testing
	time_bp_to_1st_RAR_yr	Time interval (years) between first office visit with blood pressure recorded to aldosterone/reni
	time_enc_to_1st_RAR_yr	Time interval (years) between first clinical encounter to aldosterone/renin test
Laboratory data	Test_*	
Encounter	enc_n	Number of clinical encounters
	enc_bp_n	Number of office visits with blood pressure recorded
	time_bp_after_1st_RAR_yr	Time interval (years) between aldosterone/renin test and last office visit with blood pressure
	time_enc_after_1st_RAR_yr	Time interval (years) between aldosterone/renin test and last clinical encounter
Diagnoses	Dx_*	
	CODE_*	
Notes	re_*	count of key words in clinical notes

Table C.2: Top 20 features selected at the first stage by LASSO and Adaptive LASSO respectively for the PA screening test using the UPHS EHR data.

LASSO	Adaptive LASSO
RAR_tests_n	CODE_403.11_n
time_enc_to_1st_RAR_yr	CODE_E26.09_n
time_enc_after_1st_RAR_yr	CODE_E26.81_n
CODE_E10.40_n	CODE_250.82_n
I_lab_24HR_URINE_CREATININE	CODE_E26.9_n
CODE_250.71_n	CODE_250.41_n
I_lab_IVC_Aldos	CODE_E27.9_n
re_hyperaldo_spec	CODE_327.24_n
CODE_250.93_n	CODE_250.42_n
re_salt_sensit_spec	CODE_402.00_n
re_htn	CODE_401.1_n
Test_CALCIUM_mg.dL	CODE_E11.620_n
re_htn_spec	CODE_402_n
CODE_401.1_n	CODE_250.52_n
re_adrenal_adenoma	CODE_401.0_n
Test_MAGNESIUM_mg.dL	CODE_E10.69_n
CODE_227.0_n	CODE_227.0_n
CODE_250.42_n	CODE_250.4_n
raceOTHER	CODE_I13.0_n
Dx_h0_HTN_normlized_n	CODE_405.19_n

Table C.3: Features selected by domain experts to include in the second stage for the PA phenotyping model using the UPHS EHR data.

Variables
gender
hisp
race
enc_n
enc_bp_n
dbp
sbp
high_BP_prop
Test_POTASSIUM_mmol.L
Test_SODIUM_mmol.L
Test_CARBON_DIOXIDE_mmol.L

Table C.4: Estimated accuracy measures for the PA phenotyping model derived from positive-only EHR data using proposed two-stage method when Adaptive LASSO was performed in the first stage. Both AUC were estimated at 0.82 when anchor S_1 and S_2 was used respectively.

	Anchor S_1		Anchor S_2	
Threshold	$\widehat{TPR}$	$\widehat{PPV}$	$\widehat{TPR}$	$\widehat{PPV}$
0.050	0.779	0.050	0.789	0.039
0.100	0.682	0.087	0.691	0.065
0.150	0.581	0.131	0.596	0.095
0.200	0.507	0.175	0.515	0.134
0.250	0.408	0.238	0.448	0.185
0.300	0.314	0.306	0.334	0.234
0.350	0.248	0.362	0.274	0.288
0.400	0.158	0.355	0.204	0.376
0.450	0.092	0.333	0.139	0.365
0.500	0.059	0.324	0.086	0.332

BIBLIOGRAPHY

Agarwal, V, Podchiyska, T, Banda, JM, Goel, V, Leung, TI, Minty, EP, Sweeney, TE, Gyang, E, and Shah, NH (2016). Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association* 23.6, 1166–1173.

Ananthakrishnan, AN, Cai, T, Savova, G, Cheng, SC, Chen, P, Perez, RG, Gainer, VS, Murphy, SN, Szolovits, P, Xia, Z, et al. (2013). Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflammatory bowel diseases* 19.7, 1411–1420.

Banda, JM, Halpern, Y, Sontag, D, and Shah, NH (2017). Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *AMIA Summits on Translational Science Proceedings* 2017, 48.

Bejan, CA, Xia, F, Vanderwende, L, Wurfel, MM, and Yetisgen-Yildiz, M (2012). Pneumonia identification using statistical feature selection. *Journal of the American Medical Informatics Association* 19.5, 817–823.

Carroll, RJ, Eyler, AE, and Denny, JC (2011). "Naïve electronic health record phenotype identification for rheumatoid arthritis". In: *AMIA annual symposium proceedings*. Vol. 2011. American Medical Informatics Association, 189.

Carroll, RJ, Thompson, WK, Eyler, AE, Mandelin, AM, Cai, T, Zink, RM, Pacheco, JA, Boomershine, CS, Lasko, TA, Xu, H, et al. (2012). Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association* 19.e1, e162–e169.

Christiansen, EK, Skipenes, E, Hausken, MF, Skeie, S, **O**stbye, T, and Iversen, MM (2017). Shared electronic health record systems: key legal and security challenges. *Journal of diabetes science and technology* 11.6, 1234–1239.

Denis, F, Gilleron, R, and Letouzey, F (2005). Learning from positive and unlabeled examples. *Theoretical Computer Science* 348.1, 70–83.

Ducher, M, Mounier-Véhier, C, Lantelme, P, Vaisse, B, Baguet, JP, and Fauvel, JP (2015). Reliability of a Bayesian network to predict an elevated aldosterone-to-renin ratio. *Archives of cardiovascular diseases* 108.5, 293–299.

Elkan, C and Noto, K (2008). "Learning classifiers from only positive and unlabeled data". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 213–220.

Goldstein, BA, Navar, AM, Pencina, MJ, and Ioannidis, J (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 24.1, 198–208.

Halpern, Y, Choi, Y, Horng, S, and Sontag, D (2014). "Using anchors to estimate clinical state without labeled data". In: *AMIA Annual Symposium Proceedings*. Vol. 2014. American Medical Informatics Association, 606.

Halpern, Y, Horng, S, Choi, Y, and Sontag, D (2016). Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association* 23.4, 731–740.

Hing, E and Hsiao, CJ (2014). *Percentage of Office-Based Physicians Using an Electronic Health Record (EHR) System, by State-National Ambulatory Medical Care Survey, United States, 2013*.

Hong, C, Liao, KP, and Cai, T (2019). Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. *Biometrics* 75.1, 78–89.

Hosmer, DW, Lemeshow, S, and Klar, J (1988). Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biometrical Journal* 30.8, 911–924.

Hosmer Jr, DW, Lemeshow, S, and Sturdivant, RX (2013). *Applied logistic regression*. Vol. 398. John Wiley & Sons.

Hripcsak, G and Albers, DJ (2012). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 20.1, 117–121.

Hsiao, CJ and Hing, E (2012). Use and characteristics of electronic health record systems among office-based physician practices, United States, 2001-2012. *NCHS Data Brief* 111.1, 1–8.

Jamoom, E and Yang, N (2016). Table of electronic health record adoption and use among office-based physicians in the US, by state: 2015 National Electronic Health Records Survey. *National Center for Health Statistics*.

Käyser, SC, Deinum, J, Grauw, WJ de, Schalk, BW, Bor, HJ, Lenders, JW, Schermer, TR, and Biermans, MC (2018). Prevalence of primary aldosteronism in primary care: a cross-sectional study. *Br J Gen Pract* 68.667, e114–e122.

Krenn, L and Schlossman, D (2017). Have Electronic Health Records Improved the Quality of Patient Care? *PM&R* 9, S41–S50.

Krousel-Wood, M, McCoy, AB, Ahia, C, Holt, EW, Trapani, DN, Luo, Q, Price-Haywood, EG, Thomas, EJ, Sittig, DF, and Milani, RV (2017). Implementing electronic health records (EHRs): health care provider perceptions before and after transition from a local basic EHR to a commercial comprehensive EHR. *Journal of the American Medical Informatics Association* 25.6, 618–626.

Kumar, V, Liao, K, Cheng, SC, Yu, S, Kartoun, U, Brettman, A, Gainer, V, Cagan, A, Murphy, S, Savova, G, et al. (2014). Natural language processing improves phenotypic accuracy in an electronic medical record cohort of type 2 diabetes and cardiovascular disease. *Journal of the American College of Cardiology* 63.12 Supplement, A1359.

Lazzarini, N, Nanni, L, Fantozzi, C, Pietracaprina, A, Pucci, G, Seccia, TM, and Rossi, GP (2015). Heterogeneous machine learning system for improving the diagnosis of primary aldosteronism. *Pattern Recognition Letters* 65, 124–130.

Liao, KP, Ananthakrishnan, AN, Kumar, V, Xia, Z, Cagan, A, Gainer, VS, Goryachev, S, Chen, P, Savova, GK, Agniel, D, et al. (2015). Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One* 10.8.

Liao, KP, Cai, T, Gainer, V, Goryachev, S, Zeng-treitler, Q, Raychaudhuri, S, Szolovits, P, Churchill, S, Murphy, S, Kohane, I, et al. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research* 62.8, 1120–1127.

Liu, T and Tao, D (2015). Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence* 38.3, 447–461.

Love, TJ, Cai, T, and Karlson, EW (2011). "Validation of psoriatic arthritis diagnoses in electronic medical records using natural language processing". In: *Seminars in arthritis and rheumatism*. Vol. 40. 5. Elsevier, 413–420.

Mulatero, P, Monticone, S, Burrello, J, Veglio, F, Williams, TA, and Funder, J (2016). Guidelines for primary aldosteronism: uptake by primary care physicians in Europe. *Journal of hypertension* 34.11, 2253–2257.

Oenolle, T, Chaiellier, G, Julien, J, Battaglia, C, Luo, P, and Plouin, PF (1993). Left ventricular mass and geometry before and after etiologic treatment in renovascular hypertension, aldosterone-producing adenoma, and pheochromocytoma. *American journal of hypertension* 6.11_Pt_1, 907–913.

Pathak, J, Kho, AN, and Denny, JC (2013). *Electronic health records-driven phenotyping: challenges, recent advances, and perspectives*.

Ramaswamy, H, Scott, C, and Tewari, A (2016). "Mixture proportion estimation via kernel embeddings of distributions". In: *International Conference on Machine Learning*, 2052–2060.

Romano, MJ and Stafford, RS (2011). Electronic health records and clinical decision support systems: impact on national ambulatory care quality. *Archives of internal medicine* 171.10, 897–903.

Rossi, GP, Bernini, G, Caliumi, C, Desideri, G, Fabris, B, Ferri, C, Ganzaroli, C, Giacchetti, G, Letizia, C, Maccario, M, et al. (2006). A prospective study of the prevalence of primary aldosteronism in 1,125 hypertensive patients. *Journal of the American College of Cardiology* 48.11, 2293–2300.

Rossi, GP, Rossi, E, Pavan, E, Rosati, N, Zecchel, R, Semplicini, A, Perazzoli, F, and Pessina, AC (1998). Screening for primary aldosteronism with a logistic multivariate discriminant analysis. *Clinical endocrinology* 49.6, 713–723.

Rossi, GP, Sacchetto, A, Visentin, P, Canali, C, Graniero, GR, Palatini, P, and Pessina, AC (1996). Changes in left ventricular anatomy and function in hypertension and primary aldosteronism. *Hypertension* 27.5, 1039–1045.

Shigematsu, Y, Hamada, M, Okayama, H, Hara, Y, Hayashi, Y, Kodama, K, Kohara, K, and Hiwada, K (1997). Left ventricular hypertrophy precedes other target-organ damage in primary aldosteronism. *Hypertension* 29.3, 723–727.

Shivade, C, Raghavan, P, Fosler-Lussier, E, Embi, PJ, Elhadad, N, Johnson, SB, and Lai, AM (2013). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association* 21.2, 221–230.

Song, M, Kraft, P, Joshi, AD, Barrdahl, M, and Chatterjee, N (2014). Testing calibration of risk models at extremes of disease risk. *Biostatistics* 16.1, 143–154.

Tibshirani, R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, 267–288.

Tsiatis, AA (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* 67.1, 250–251.

Wachtel, H, Zaheer, S, Shah, PK, Trerotola, SO, Karakousis, GC, Roses, RE, Cohen, DL, and Fraker, DL (2016). Role of adrenal vein sampling in primary aldosteronism: Impact of imaging, localization, and age. *Journal of surgical oncology* 113.5, 532–537.

Wang, L, Guo, Z, Ding, X, Herman, D, and Chen, J (2020). Predicting patients' disease risk by estimating the screening process with electronic health records data. *In preparation*.

Ward, G, Hastie, T, Barry, S, Elith, J, and Leathwick, JR (2009). Presence-only data and the EM algorithm. *Biometrics* 65.2, 554–563.

Weiskopf, NG and Weng, C (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* 20.1, 144–151.

Windmeijer, FAG (1990). The asymptotic distribution of the sum of weighted squared residuals in binary choice models. *Statistica Neerlandica* 44.2, 69–78.

Xia, Z, Secor, E, Chibnik, LB, Bove, RM, Cheng, S, Chitnis, T, Cagan, A, Gainer, VS, Chen, PJ, Liao, KP, et al. (2013). Modeling disease severity in multiple sclerosis using electronic health records. *PloS one* 8.11.

Yadav, P, Steinbach, M, Kumar, V, and Simon, G (2018). Mining electronic health records (EHRs): a survey. *ACM Computing Surveys (CSUR)* 50.6, 85.

Yu, S, Chakrabortty, A, Liao, KP, Cai, T, Ananthakrishnan, AN, Gainer, VS, Churchill, SE, Szolovits, P, Murphy, SN, Kohane, IS, et al. (2016). Surrogate-assisted feature extraction for high-throughput phenotyping. *Journal of the American Medical Informatics Association* 24.e1, e143–e149.

Yu, S, Chakrabortty, A, Liao, KP, Cai, T, Ananthakrishnan, AN, Gainer, VS, Churchill, SE, Szolovits, P, Murphy, SN, Kohane, IS, et al. (2017). Surrogate-assisted feature extraction for high-throughput phenotyping. *Journal of the American Medical Informatics Association* 24.e1, e143–e149.

Zhang, L, Ding, X, Ma, Y, Muthu, N, Ajmal, I, Moore, J, Herman, D, and Chen, J (2019). A maximum likelihood approach for Electronic Health Record phenotyping using positive and unlabeled patients. *Journal of American Medical Informatics Association, Accepted*.

Zhang, L, Ma, Y, Herman, D, and Chen, J (2020). Testing calibration of risk prediction models using positive-only electronic health record data. *Biostatistics, submitted*.

Zou, H (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101.476, 1418–1429.