Multivariate Inference and Surveillance using Population Scale Data

Xu Shi

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Patrick Heagerty, Chair

Andrea Cook, Chair

Marco Carone

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Multivariate Inference and Surveillance using Population Scale Data

Xu Shi

Co-Chairs of the Supervisory Committee:
Professor Patrick Heagerty
Department of Biostatistics

Professor Andrea Cook
Department of Biostatistics, Kaiser Permanente Washington Health Research Institute

Recent federal initiatives are incentivizing the routine collection and linkage of electronic health records (EHR) data across clinics, hospitals, and healthcare systems. Use of large-scale EHR data presents numerous opportunities for biomedical research but also unique challenges as EHR data are not collected for research purposes. We first consider data quality issues and the learning of differential patterns in healthcare utilization, through multivariate testing and estimation of subgroup differences in the endorsements of billing codes. We further consider the critical problem of pharmacosurveillance to monitor for rare adverse events once a drug or product is incorporated into routine clinical care. Key issues are the need to provide formal statistical inference for rare outcomes, and to offer flexible methods to control for many potential confounders. We provide an influence function based statistical framework that incorporates recent theoretical advances from econometrics to study conditions under which a three-step approach using regression adjustment of propensity score would provide valid and efficient estimation. The influence function representation also provides a variance estimator that fully accounts for uncertainty of both outcome modeling and propensity score estimation. We finally consider the potential correlation within healthcare providers and a simple correction in variance estimation for valid inference.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my sincerest appreciation to my advisors, Patrick Heagerty and Andrea Cook, for their implicit and explicit help, and for their professional and personal impact on me. I would also like to thank my dissertation committee members Marco Carone, and Josh Carlson. I am deeply grateful to Marco Carone for introducing me to the research area of modern inference in semiparametric and nonparametric models. Finally, I would like to thank my friends and my family for their love and support.

DEDICATION

To my parents, Yuhui and Erqin.

Chapter 1
INTRODUCTION

The growing availability of EHR data has opened up new opportunities to evaluate factors associated with both individual benefit and potential harm. For example, the FDA Sentinel Initiative recently created a surveillance network with over 100 million patient lives to monitor the safety of all new medical products. Often pre-approval trials are too small to systematically detect rare adverse events and may not be generalizable to the population who will ultimately receive the medical products post-approval. Therefore, the FDA created Sentinel Initiative using prospective observational cohorts of administrative EHR databases from numerous large health plans across the US. Similarly, PCORI recently created PCORnet which links EHRs from geographic regions throughout the US with the ultimate goal of facilitating faster, easier, and less costly comparative effectiveness research. However, EHR data are not collected for research purposes. Such cost-effective data sources come with unique challenges, which motivated the methodological research in this dissertation.

- **High-throughput Comparison of Healthcare Utilization using EHR Data**

Recent federal initiatives are incentivizing the routine collection and linkage of EHR data across clinics, hospitals, and healthcare systems. One key challenge to the use of electronically assembled cohorts is the potential for variation in both the choice of specific procedures and coding practices across healthcare systems. This variation could be due to systematic differences in patient populations, and may also be linked to data quality issues.

We develop statistical methods for evaluation of data quality and comparison of healthcare utilization between select patient subgroups. We focus on the Current Procedural Terminology (CPT) codes which are used in a standardized fashion to record patient treat-

ment histories, but the methods we develop can be used for any structured EHR data such as diagnostic codes (ICD-9 or 10). We specifically study testing procedures that are valid for both rare and common codes as routinely encountered with medical procedures and we also transfer methods from genetic association studies. Hierarchical structure in terms of both thematically grouped medical codes and provider-level clustering adds unique complexity to the analysis of EHR data. We consider penalized regression methods unifying estimation and inference with hierarchical shrinkage to leverage such structure and stabilize estimates for rare procedures. We also expand inference methods to account for potential correlation driven by provider behavior. We ultimately provide interpretable dynamic graphical tools that can help researchers to explore the healthcare utilization patterns, implemented in both an R package and an interactive web application.

- **Flexible Confounding Adjustment Targeting Population-level Inference**

Although EHR data provide a large sample size and rich patient information from a broad population, when conducting safety surveillance, the outcome of interest, such as myocardial infarction (MI), is extremely rare. Conventional covariate adjustment methods may have convergence issues when the outcome is rare and flexible nonparametric regression may be challenging due to the high-dimensionality of the covariates. In contrast, given sufficient uptake of both exposure and comparator medical products, methods that use propensity score models, which estimate the probability of exposure given confounders, may be more attractive.

A key idea is to substitute the set of all covariates with a propensity score in the outcome regression model, which provides both sufficient control of confounding and dimension reduction. We characterize nonparametric propensity score adjustment in a generalized varying coefficient model to introduce flexibility and reduce model assumptions. To estimate marginal, population-level contrasts that are the central focus of causal inference, the use of a regression model is often considered as the intermediate summary that is then used in

a final standardization step, which takes the empirical average of the pair of predicted risks over the entire population under hypothetical exposure and control.

Flexible adjustment of the propensity score in an outcome regression model has been proposed but no formal representation of the statistical procedure has been detailed. We provide an influence function based statistical framework that incorporates recent theoretical advances from econometrics that permit the study of conditions under which the three-step approach (propensity score estimation, flexible outcome regression, and standardization) would provide valid and efficient estimation. In addition, the influence function representation provides a direct and simple variance estimator that fully accounts for uncertainty of both outcome modeling and propensity score estimation. We conduct a realistic simulation study to evaluate our proposed strategy and compare this three-step approach to common alternative methods, mimicking real data from a recent FDA Sentinel investigation that compares the effect of angiotensin-converting enzyme inhibitors and beta blockers on incidence of angioedema in the first 30 days.

- **Valid Inference under Provider-level Clustering**

There is growing attention to the issue of lack of statistical independence in electronic health record (EHR)-based clinical research, which collects data at the level of aggregates of healthcare providers. In particular, patients treated by the same provider are more likely to receive similar treatments and respond in a similar manner. As such, patient outcomes may be clustered within a provider, which adds unique complexity to the analysis of EHR data and may lead to incorrect inference if not appropriately accounted for. We consider a simple correction of the influence function based variance estimator to account for potential clustering among patients treated by the same provider. We also evaluate the impact of ignoring provider-level clustering and the performance of the variance correction through an extension of the realistic simulation study.

Chapter 2

COMPARING HEALTHCARE UTILIZATION PATTERNS VIA GLOBAL DIFFERENCES IN THE ENDORSEMENT OF CURRENT PROCEDURAL TERMINOLOGY CODES

This work is in press in Annals of Applied Statistics [95].

2.1 Introduction

In the United States, the use of electronic medical records (EMR) is now incentivized due to the 2009 enactment of the Health Information Technology for Economic and Clinical Health (HITECH) Act. Large scale EMR data will open new opportunities for research to improve patient care and the health of the public. Current national research efforts include linking EMR to conduct pharmacosurveillance (e.g., FDA Sentinel) and assembling large clinical populations for comparative effectiveness research (e.g., PCORnet). One key element of EMR data is the recording of patient treatment history via the Current Procedural Terminology (CPT) coding system. CPT codes are five-character codes describing medical services and procedures, and are used for patient management and billing. CPT codes provide a standardized description that allows communication across providers and systems, and facilitate identification of clinical information for comparative effectiveness research.

A natural research question is whether different subgroups of patients have different healthcare utilization patterns, and interest may lie in the entire spectrum of all potential services. A motivating example is the Back pain Outcomes using Longitudinal Data (BOLD) project, which enrolled 5239 patients who are 65 years of age or older with back pain from multiple healthcare systems with a primary interest in early imaging [50]. In this study we combine electronic health records across sites, and there is need to assess poten-

tial data quality concerns by comparing codes between healthcare systems. Ultimately the BOLD study compared healthcare utilization for propensity score matched patients with and without early radiologic imaging. We were interested in the effect of early imaging on down stream healthcare utilization in terms of a summary measure of total spine-related procedures, but we also wanted to examine the full set of individual CPT codes and overall utilization. There is an emerging need to develop inference methods that are tailored for the electronic medical records context. Recent federal healthcare and research initiatives are incentivizing the routine collection of population scale data, and statistical methods are needed for evaluation of data quality and for high-throughput comparison of utilization for select patient subsets.

However, the use of EMR data for research comes with several challenges. First, a unique aspect of EMR data is the organizational structure where individual patients are typically nested within providers who may have unexplained variation in their treatment patterns that induce correlation in utilization indicators for their patient panel. Second, an important characteristic of CPT codes is the hierarchical structure in coding taxonomy: multiple CPT codes may represent similar or related procedures. According to the Clinical Classifications Software (CCS) for Services and Procedures, CPT codes are naturally collapsible into 244 clinically meaningful groups, which define major categories of procedures. The CCS-Services and Procedures taxonomy is a part of the Healthcare Cost and Utilization Project (HCUP), a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality. See `https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp` for details on the CCS classification. Inference can be made at either the code or the group of codes level, and shrinkage methods may be desired to borrow strength from similar codes. Third, in practice, both investigation of data quality and evaluation of overall utilization imply the need to perform inference for thousands of codes, and robust methods that can be applied to both common codes and rare procedures are needed. In this chapter, we consider testing and estimation methods for quantifying the significance and magnitude of differences in the delivery of all possible procedures and services between

two cohorts of patients. Our proposed methods are tailored to the unique and increasingly important context of healthcare delivery system generated observational data with the above mentioned challenges: patient clustering, CPT code grouping, and coexistence of common and rare procedures.

This chapter is organized as follows. Section 2.2 details testing for differences in healthcare utilization. Section 2.2.1 focuses on code-wise testing procedures for both common and rare codes. Section 2.2.2 considers testing of procedures defined by a group of codes. Section 2.2.3 discusses methods that account for provider-level clustering. In Section 2.3 we propose rate ratio estimation and inference methods. Section 2.3.1 details a ridge regression model which takes advantage of the hierarchical structure of CPT codes and stabilizes estimates of rare procedures. In Section 2.3.2 we provide inference method accounting for shrinkage bias. Section 2.3.4 considers inference with provider-level clustering. Both methods in Sections 2.2 and 2.3 allow for confounding adjustment. In Section 2.4 we conduct simulation studies to evaluate the performance of the code-wise and group-wise tests, as well as the inference for ridge regression. We also study the influence of provider-level clustering on testing procedures and the performance of methods that accounts for within-provider correlation. In Section 2.5 we illustrate the methods by analyzing EMR data from the BOLD study. We evaluate differences in CPT codes assigned among patients from two healthcare systems: Kaiser Permanente in Northern California, and Henry Ford Health System in Detroit. For healthcare systems or clinics within systems, the benchmarking of one site against a reference site is an important part of revealing variation that may require attention in order to align delivery decisions with clinical guidelines or to potentially reduce cost. Therefore, differences in healthcare utilization across the entire set of codes are generally important to evaluate for both delivery assessment and research purposes. We develop and illustrate graphical tools that compare patient subgroups across the full spectrum of procedures and services for exploratory research using large scale healthcare data. We close with a discussion in Section 2.6.

2.2 Testing for Utilization of Procedures Defined by a CPT Code or a Block of Codes

For simplicity, we consider patients' visits over a fixed time period such as one year, although methods that characterize rates of code endorsement can easily handle variable follow-up time at the patient level by weighting the outcome with the inverse of patient-specific length of follow-up. Let n_s, $s = 0, 1$, denote the total subjects in cohort s, and without loss of generality, we assume one year of follow-up for each subject. For a specific procedure described by CPT code c, we take two approaches to comparing utilization rates across cohorts. First, we consider an outcome based on a count of how many times the procedure was delivered to patient i in cohort s over the year, denoted as Y_{si}^c, where $s = 0, 1$, $i = 1, \ldots, n_s$, $c = 1, \ldots, C$. Second, for certain scientific questions we may only be interested in whether the procedure was ever endorsed for a subject, and we may choose to dichotomize the count data into any/none outcomes $Z_{si}^c = \mathbb{1}(Y_{si}^c > 0)$, as an indicator of whether patient i was assigned code c in any visits over the year.

2.2.1 Testing for Code-specific Patient Utilization

Our interest in CPT code-wise inference requires selection of a testing strategy that can be valid for both common and rare codes, and under potential overdispersion. However, in practice, there is little guidance on how to choose appropriate tests. In this section we discuss various two-sample testing strategies that are candidates for the evaluation of variation in code endorsement rates across cohorts for count and binary data with a goal of characterizing the applied options, summarized in Table 2.1. In Section 2.4 we perform numerical studies to illustrate the performance of testing options for a range of rate parameters that may be expected from CPT data. Although we focus on a crude comparison, adjustment for covariates can be achieved through stratification or matching on the propensity score, which is the probability that a patient belongs to a cohort given the observed confounders [87, 88].

Table 2.1: Summary of two-sample testing options for CPT count and binary Data. The notation Λ denotes the likelihood ratio test statistic.

Distribution		**Likelihood Ratio Test**	**Conditional Exact Test**
Count	**Poisson**	$Y_{si}^c \sim \text{Poisson}(\lambda_s^c)$ $-2\log(\Lambda) \overset{H_0}{\to} \chi^2$	$Y_1^c \mid Y_0^c + Y_1^c \overset{H_0}{\sim} \text{Binomial}($ $n = Y_0^c + Y_1^c,\ p = \frac{n_1}{n_0+n_1})$
	Negative Binomial	$Y_{si}^c \sim \text{Neg-Bin}(p = \frac{\lambda_s^c}{\lambda_s^c+\frac{1}{\phi^c}}, r = \frac{1}{\phi^c})$ $-2\log(\Lambda) \overset{H_0}{\to} \chi^2$	$\Pr(Y_0^c = y_0^c, Y_1^c = y_1^c \mid Y_0^c + Y_1^c) \overset{H_0}{=}$ $\binom{y_0^c+\frac{n_0}{\phi^c}-1}{y_0^c} \cdot \binom{y_1^c+\frac{n_1}{\phi^c}-1}{y_1^c} \Big/ \binom{y_0^c+y_1^c+\frac{n_0+n_0}{\phi^c}-1}{y_0^c+y_1^c}$
Binary	**Binomial**	$Z_{si}^c \sim \text{Bernoulli}(p_s^c)$ $-2\log(\Lambda) \overset{H_0}{\to} \chi^2$	$Z_1^c \mid Z_0^c + Z_1^c \overset{H_0}{\sim} \text{Hypergeometric}($ $N = n_0 + n_1, n = Z_0^c + Z_1^c, K = Z_1^c)$
Semi-parametric		t-test	

2.2.1.1 Count Outcome

For count data, a natural model is the Poisson distribution characterized by a rate parameter, λ_s^c, with $Y_{si}^c \sim \text{Poisson}(\lambda_s^c)$, $i = 1, \ldots, n_s$. In cohorts where there are both relatively healthy and extremely ill patients, there will be overdispersion, and the simple Poisson mean-variance relationship will not hold. In this situation, use of negative binomial distribution provides one model-based generalization of the Poisson assumption. The negative binomial model contains a rate parameter and an additional parameter that characterizes overdispersion: $Y_{si}^c \sim$ negative binomial$(p = p_s^c = \frac{\lambda_s^c}{\lambda_s^c+\frac{1}{\phi^c}}, r = 1/\phi^c)$, $i = 1, \ldots, n_s$, where ϕ^c is the code-specific overdispersion parameter which is shared by patients across cohorts. The mean and variance are parameterized as $\text{E}[Y_{si}^c] = \lambda_s^c$ and $\text{Var}[Y_{si}^c] = \lambda_s^c(1 + \lambda_s^c\phi^c)$. We wish to test whether code endorsement varies by cohort, i.e., H_0: $\lambda_0^c = \lambda_1^c$.

In the EMR setting, the number of patients is usually quite large (thousands or greater) and large sample approximations should be valid. Therefore, if the model assumption was valid, then a likelihood ratio test (LRT) using observations from each patient could provide inference regarding coding rates across cohorts. However, rare codes can lead to low expected cell counts, which can lead to a poor χ^2 approximation for the null distribution of the LRT.

In this case, as an exact alternative, we can collapse patient-level information within a cohort by computing the total counts $Y_s^c = \sum_{i=1}^{n_s} Y_{si}^c$ and apply a conditional exact test (ET) which calculates a conditional probability that does not depend on large sample approximations [78, 85].

While use of the negative binomial model allows a partial decoupling of the mean and the variance, it may not provide valid inference when the true data generating mechanism is not adequately characterized by a simple overdispersed count model. Alternatively, with large sample sizes and any underlying distribution, we can use the two-sample t-test as a semi-parametric method for testing.

2.2.1.2 Any/None Outcome

Use of endorsement rates will often be the appropriate strategy for answering scientific questions about variation in utilization. However, for certain codes such as recommended annual screening measures or vaccinations, it may be desirable to simply analyze the count data as a derived binary outcome since the clinical significance is indicated by any endorsement of the code. An any/none outcome indicates whether a patient was ever assigned a code during his or her visits over the year. It can be modeled as $Z_{si}^c \sim \text{Bernoulli}(p_s^c)$ for a patient, and $Z_s^c = \sum_{i=1}^{n_s} Z_{si}^c \sim \text{Binomial}(n_s, p_s^c)$ for a cohort. Our goal is to test whether the probability of assigning a CPT code varies by cohort, that is, H_0: $p_0^c = p_1^c$. When the requirements of the χ^2 approximation are met for the LRT, we use the Binomial LRT. When the expected cell counts are too small, we can use the conditional ET assuming Binomial model, which is the well-known Fisher's ET for a two-by-two table constructed using cohort level data.

We summarize standard asymptotic LRTs and ETs for count and binary data in Table 2.1, and provide detailed reviews in Appendix A.1. In summary, key practical issues include: whether to adopt asymptotic or exact tests; whether to consider a count or indicator outcome; and whether or how to account for overdispersion in CPT counts.

2.2.2 Testing for Block-specific Patient Utilization

It has been shown that for a specific procedure, the level of agreement among coders and agencies in assigning CPT codes can be poor [51, 52, 12, 48]. That is, physicians might use different codes to describe the same procedure since multiple codes can be appropriate for a certain general procedure. For example, bilateral screening mammography, according to the imaging technology, can be coded using CPT code 77057 which is labelled "Screening mammography, bilateral (2-view film study of each breast)", or can be recorded using CPT code G0202 labelled as "Screening mammography, producing direct digital image, bilateral, all views". Therefore, code-level analysis may detect variation that is not reflective of meaningful practice variation, and analysis at a "code group" level may be more appropriate. According to the CCS, CPT codes can be collapsed into groups. We call such a group of codes a "block". Procedures can be compared at CCS block-level where finer scale differences in procedure coding may not indicate an overall difference, so that the comparison is not sensitive to physicians' choice of codes within a block.

Suppose that there are C total codes that can be categorized into B blocks. Let $S(b)$ be the set of codes that belong to block b, and let C_b denote the number of codes in $S(b)$, such that $\sum_{b=1}^{B} C_b = C$. Given the hierarchical structure of medical codes, we use $Y_{si}^{bc} = Y_{si}^{c}$ to emphasize that each code c belongs to a certain block b. Therefore, the count vector Y_{si}^{bc} for $b = 1, \ldots, B$ and $c = 1, \ldots, C_b$ corresponds to one observation or row of data associated with patient i in cohort s. In the following sections we detail testing methods that can be used to make inference at the block level.

2.2.2.1 Burden Test

A simple testing procedure parallels methods used for genomic data that have been termed "burden tests", since the total number of endorsements within a block (i.e., total burden) is used as the basis for testing [58, 65, 66]. Using this approach we apply the code-wise testing methods in Section 2.2.1 to block-level summaries. Specifically, for a procedure defined by

block b, $Y_{si}^b = \sum_{c \in S(b)} Y_{si}^{bc}$ summarizes the assignment of the block-specific procedure to patient i in cohort s over the year, and $Z_{si}^b = \mathbb{1}(Y_{si}^b > 0)$ describes whether the procedure was assigned to patient i at any visits over the year. Within genomic research, the burden test is a group-wise association test that potentially increases power by combining genetic counts within regions or genes [107, 54]. In our context, when codes within a given block are consistently used more frequently in one cohort, the burden test accumulates individual code effects to increase power. Such a strategy is especially important when dealing with rare codes. However, for codes that have inconsistent distributions, combining their effects when they may be in opposite directions across the comparison groups can lead to cancellation of potentially meaningful variation and null test results. Thus, the burden test might diminish code-level effects when they are aggregated. On one hand, such aggregation ensures that the burden test is insensitive to code substitution, but on the other hand, it may decrease power to detect a meaningful code-level variation.

2.2.2.2 Sequence Kernel Association Test

We use the sequence kernel association test (SKAT) as a complement to the burden test [107]. This test is based on a mixed model framework that was developed to collectively test for the association between a set of genetic variants and a phenotype. We treat CPT codes within a block as analogues to genetic variants within a region, and the cohort of a patient as the dichotomous phenotype to allow testing for groups of procedure codes.

For each block b, we consider the logistic regression model for the phenotype, i.e., for cohort s

$$\text{logit}(\Pr(s_i = 1 \mid Y_{si}^{bc})) = \alpha_0 + \sum_{c \in S(b)} \beta_c Y_{si}^{bc} \ ,$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{C_b}) \sim N(\mathbf{0}, \sigma^2 \boldsymbol{W})$, where $\boldsymbol{W} = \text{diag}(w_1, ..., w_{C_b})$ is a weight matrix. Testing for the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is equivalent to the variance component test for $H_0 : \sigma = 0$. Let $\hat{\mu}_0$ denote the expectation of outcome $\Pr(s = 1)$ under the null. The score test statistic is $\mathbf{Q} = (\mathbf{s} - \hat{\mu}_0 \mathbf{1})^T \mathbf{K} (\mathbf{s} - \hat{\mu}_0 \mathbf{1})$, where $\mathbf{K} = \mathbf{Y} \mathbf{W} \mathbf{Y}^T$ is a weighted

kernel, and $\mathbf{Y}$ is an $(n_0 + n_1) \times C_b$ matrix whose elements are Y_{si}^{bc}. It is important to note that SKAT can be applied to either a matrix of binary data $\mathbb{1}(Y_{si}^{bc} > 0)$ similar to single nucleotide polymorphism (SNP) data, or count data Y_{si}^{bc} similar to RNA-seq data. In addition, adjustment for confounders can be achieved through fitting the model

$$\text{logit}(\Pr(s = 1)) = \alpha_0 + \alpha Z_i + \sum_{c \in S(b)} \beta_c Y_{si}^{bc} ,$$

where Z_i denotes the set of covariates for patient i [107].

SKAT collectively tests for the association between the endorsement of codes within a block and the cohort. It can increase power by summarizing over multiple codes but does not require that associations are all in the same direction. However, when the code effects are truly in the same direction, simulation studies showed that burden tests may have higher power [107]. Thus, we recommend that both the burden test and the SKAT are used with awareness of the types of departures that would likely be detected.

2.2.3 Provider-level Clustering

In addition to patient-level variation, there may be provider-level variation in use of CPT codes. Specifically, providers may have individual preferences in their typical choice of treatment paths, which ultimately introduces correlation between patients treated by the same provider.

The impact of ignoring correlation for clusters of patients generally lies in estimation of the standard errors or test statistic variances. For two-sample testing procedures employed in code-wise comparisons and block-wise burden test, we consider using a generalized estimating equation (GEE) type sandwich variance estimator with working independence covariance matrix in a z-statistic to replace the t-test [26]. For SKAT, [79] considered expanding SKAT for longitudinal data and developed the longitudinal kernel machine regression (L-KM) method. In our context, the L-KM essentially adds a provider-level random intercept to introduce correlation between patients within provider. Compared to SKAT, the variance-covariance matrix in the test statistic is estimated from a mixed model which takes into

account correlation within provider.

When the primary care provider is taken to define the cluster, patients are nested within providers and our proposed testing procedure with a GEE type sandwich estimator can correct the type I error to validly account for any provider level clustering. However, when patients are treated by potentially different providers, there could be multilevel clustering that is not necessarily nested. In this case, variance estimation accounting for such multilevel clustering requires extra attention and additional complexity. One strategy would be a likelihood-based random effects approach with implementation using Bayesian computational techniques to provide valid inference. Alternatively, the GEE type variance estimator with working independence covariance matrix proposed in our testing procedures can be generalized to account for non-nested clustering, which was introduced in [62] and [63]. The non-nested sandwich variance calculation is surprisingly simple and easy to generalize to more than two non-nested levels of clustering.

In Section 2.4.1.2, we conduct simulation studies to investigate the sensitivity/robustness to provider-level clustering of the two-sample tests. We also study the performance of the above tests with appropriate correction for within-provider clustering.

2.3 Rate Ratio Estimation and Inference using Ridge Regression

In Sections 2.2.1 and 2.2.2 we detailed testing methods that assess the statistical significance associated with the observed difference in healthcare utilization across two cohorts. However, testing provides only a partial characterization of utilization differences, and two further questions are of interest: which cohort is the frequent user, and how large is the magnitude of difference? Recall that CPT codes may be rarely used (e.g., zero counts can be common) and are nested within blocks. We take advantage of such a hierarchical structure and stabilize estimates of rare codes by using a Poisson regression model (Equation 2.1) with Ridge penalty [47] that estimates code-specific cohort effects for all codes simultaneously, and exploits potential similarity of endorsement trends within blocks.

2.3.1 Rate Ratio Estimation using Hierarchical Shrinkage

Let $Y_{s'i'}^{b'c'}$ denote the number of assignments of code c' within block b' to patient i' in cohort s'. Recall that in Section 2.2.2 we let $Y_{\cdot}^{bc} = Y_{\cdot}^{c}$ to emphasize that code c belongs to a certain block b. For each patient i' and for each code c' assigned to this patient, define indicators that denote $(s',\ b',\ c')_{i'}$ using $\text{cohort}_{1,i'} = \mathbb{1}\{s' = 1\}$, $\text{block}_{b,i'} = \mathbb{1}\{b' = b\}$, and $\text{code}_{c,i'} = \mathbb{1}\{c' = c\}$, where $b = 2, \ldots, B$ and $c \in \{S(b) \setminus c_{ref}^{b}\}$, with reference levels $s = 0$ for cohorts, $b = 1$ for blocks, and $c = c_{ref}^{b}$ for codes in block b. Take $t_{i'}$ as the offset to account for potentially different lengths of follow-up across patients, which also defines the rate of code endorsement for patient i' as $\mathrm{E}[Y_{s'i'}^{b'c'}]/t_{i'}$. In addition, we consider confounding adjustment and denote the vector of observed covariates for patient i' as $\boldsymbol{Z}_{i'}$. The model is

$$\begin{aligned}\log[E(Y_{s'}^{b'c'})] = \log(t) + \alpha_0 + \alpha_1 \text{cohort}_1 + \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\theta} \\ + \sum_{b=2}^{B} \alpha_b \text{block}_b + \sum_{b=2}^{B} \gamma_{b1} \text{block}_b \cdot \text{cohort}_1 \\ + \sum_{b=2}^{B} \sum_{c \in \{S(b) \setminus c_{ref}^{b}\}} \eta_c \text{code}_c + \sum_{b=2}^{B} \sum_{c \in \{S(b) \setminus c_{ref}^{b}\}} \zeta_{c1} \text{code}_c \cdot \text{cohort}_1\end{aligned} \tag{2.1}$$

where we suppress the notation i' for simplicity.

In this model, the rate ratio comparing cohorts 1 and 0 is determined by three components: the main effect of cohort, α_1; the block-cohort interaction that adds an increment to the overall cohort level, γ_{b1}; and the code-cohort interaction that adds an increment to the overall block level, ζ_{c1}. In other words, the rate ratio on the log scale is defined as

$$\log(\text{RR}) = \alpha_1 + \gamma_{b1} + \zeta_{c1}.$$

The rationale for building a multi-level structure that includes all blocks and codes is to leverage hierarchical shrinkage to control the extent of information sharing and to stabilize rate ratio estimates for rare procedures, with a primary goal of visualization of the effect sizes. A ridge penalty governs the shrinkage of code-specific rate ratio estimates toward the block-level rate ratio, which essentially represents an average over all codes within the same block.

In this way, we allow rare codes to borrow information from similar codes within the blocks. We also penalize across blocks to provide a second level of shrinkage for any set of codes that may also be rarely used. Therefore, we exploit the hierarchical taxonomy of procedure codes and employ two stages of penalization. Note that such hierarchical increment can be generalized to introduce nested blocks by including one indicator for each (sub-)block level. For example, CPT codes 10000 - 69990 belong to a block denoting surgery. Within this block, there are several sub-blocks denoting general surgery (10000 - 10022), integumentary system (10040 - 19499), and etc. Information for estimating coefficient of a certain block level comes from utilization of all codes and sub-blocks within this level.

The estimation of ridge regression is to minimize the negative log-likelihood plus a penalty function

$$P(\lambda_{\text{ridge}}, \omega_0, \omega_1, \omega_2) = \lambda_{\text{ridge}} \cdot \Big[\omega_0\big(\|\alpha_0\|_2^2 + \|\alpha_1\|_2^2 + \|\boldsymbol{\theta}\|_2^2\big) + \omega_1\big(\|\boldsymbol{\alpha_b}\|_2^2 + \|\boldsymbol{\gamma_{b1}}\|_2^2\big) + \omega_2\big(\|\boldsymbol{\eta_c}\|_2^2 + \|\boldsymbol{\zeta_{c1}}\|_2^2\big)\Big].$$

The form of the penalties on the coefficients guides the properties of the model. The tuning parameter λ_{ridge} controls the strength of the penalty, and $\omega_0, \omega_1, \omega_2 \in [0, 1]$ allow varying penalties to different coefficients. In particular, $\lambda_{\text{ridge}}\omega_1$ controls variation in the block effects, and $\lambda_{\text{ridge}}\omega_2$ controls shrinkage of code effects toward their overall block effect. The shrinkage is particularly important for rare codes, which could yield extreme crude estimates.

A caveat is that we need to choose the value of λ_{ridge} and ω. Throughout this work we introduce hierarchical shrinkage by fixing $\omega_0 = 0$, $\omega_1 = 0.5$, and $\omega_2 = 1$, which puts no penalty on the cohort, and restricts the penalization of block such that it is half as strong as the penalization of code. For λ_{ridge}, a sequence of 100 values is calculated corresponding to the regularization path. Because our primary goal with penalization is to simply stabilize rate ratios for codes with sparse data and/or zero counts, we choose the 95th λ_{ridge} along the sequence which gives a model with small to moderate penalization.

2.3.2 Inference for Ridge Estimation in Poisson Model

Applying the ridge penalty has the effect of shrinking the estimates toward zero, which introduces bias but reduces the variance and stabilizes the estimates [47]. Testing for potentially sparse and overdispersed CPT code utilization data can benefit from this property as long as we could de-bias the estimate in order to perform valid inference. Here we introduce a testing procedure for ridge regression in the Poisson family with log-link, referred to as the ridge test hereafter. We note that the proposed method can be adapted to any generalized linear model with a canonical link.

We simplify the notation and rewrite the model as

$$\log[E(\boldsymbol{Y} \mid \mathbf{X})] = \log(\boldsymbol{t}) + \mathbf{X}\boldsymbol{\beta}, \tag{2.2}$$

where $\boldsymbol{Y}$ is the number of assignments for all patients and all codes, $\boldsymbol{t}$ is the offset denoting lengths of follow-up, $\mathbf{X}$ is the design matrix containing the intercept, indicator cohort_1, covariates $\mathbf{Z}$, indicators for all blocks: block_b, $\text{block}_b \cdot \text{cohort}_1$, and indicators for all codes: code_c, $\text{code}_c \cdot \text{cohort}_1$, and coefficient $\boldsymbol{\beta} = \left[\alpha_0, \alpha_1, \boldsymbol{\theta}^T, \boldsymbol{\alpha_b}^T, \boldsymbol{\gamma_{b1}}^T, \boldsymbol{\eta_c}^T, \boldsymbol{\zeta_{c1}}^T\right]^T$. The penalty function can be written as $P(\cdot) = \boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta}$, where $\boldsymbol{\Lambda} = \lambda_{\text{ridge}} \cdot \boldsymbol{W}$, with penalty weight matrix $\boldsymbol{W} = \text{diag}\{\omega_0 \cdot \mathbf{1_0},\ \omega_1 \cdot \mathbf{1_1},\ \omega_2 \cdot \mathbf{1_2}\}$.

Lemma 2.1. *We assume that we are in the common EMR situation where $n > p$ since the population sizes under study are commonly large and the number of observations in the Poisson regression is driven by both the number of patients and the dimension of the multivariate outcome. Let $\hat{\boldsymbol{\beta}}_\lambda$ be the ridge shrinkage estimator with regularization parameter $\lambda_{ridge} > 0$. Then, the debiased estimator correcting for shrinkage bias is*

$$\hat{\boldsymbol{\beta}}_{\text{debias}} = \left\{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Lambda}]^{-1}\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)\right\}^{-1}\hat{\boldsymbol{\beta}}_\lambda,$$

where $\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) = \mathbf{X}^T\text{diag}(e^{\log(\boldsymbol{t})+\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda})\mathbf{X}/n$. Let $\boldsymbol{\beta}^*$ be the population coefficient satisfying model (2.2), and let $\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}$ be the population ridge coefficient from model (2.2) with penalty function $P(\cdot) = \boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta}$ and $\boldsymbol{\Lambda} = \lambda_{\text{ridge}} \cdot \boldsymbol{W}$. Then, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{debias} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0, [\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})]^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}^*)[\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})]^{-1}).$$

where $\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}) = \mathrm{E}[\boldsymbol{X}\mathrm{diag}(e^{\log(t)+\boldsymbol{X}^T\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}})\boldsymbol{X}^T]$ *and* $\boldsymbol{\Omega}(\boldsymbol{\beta}^*) = \mathrm{Var}[\boldsymbol{X}(Y - e^{\log(t)+\boldsymbol{X}^T\boldsymbol{\beta}^*})]$.

A proof is detailed in Appendix A.6. Based on Lemma 2.1, a z-statistic for calculating a p-value is

$$\Big(\mathrm{diag}\big\{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\hat{\boldsymbol{\Omega}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{debias}})[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\big\}\Big)^{-\frac{1}{2}}\sqrt{n}\hat{\boldsymbol{\beta}}_{\mathrm{debias}},$$

where $\hat{\boldsymbol{\Omega}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{debias}}) = \mathbf{X}^T\big\{\mathrm{diag}[(\boldsymbol{Y} - e^{\log(\boldsymbol{t})+\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{debias}}})^2]\big\}\mathbf{X}$ is a sandwich estimator of the variance $\boldsymbol{\Omega}(\boldsymbol{\beta}^*)$. The confidence interval is

$$\hat{\boldsymbol{\beta}}_{\mathrm{debias}} \pm Z_{1-\alpha/2}\Big(\mathrm{diag}\big\{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\hat{\boldsymbol{\Omega}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{debias}})[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\big\}\Big)^{\frac{1}{2}}/\sqrt{n}.$$

Significance testing for regularized regression is a contemporary topic in the statistical literature, and there are very few publications addressing testing with ridge regression. [15] developed a testing procedure for ridge regression in the high-dimensional setting in which the number of regression parameters p is larger than the sample size n, and assuming a deterministic design matrix. Similar to their work, we now account for shrinkage bias that results from penalization and derive the distribution of a debiased estimate. However, we took a slightly different path: we debiased by appropriately rescaling the penalized estimator instead of subtracting an estimated bias term. The latter approach requires an initial consistent estimator of the true coefficient. In addition, our testing method tackles a problem that is different from the setting in [15]. Although the number of CPT codes is large, our problem remains essentially a low-dimensional problem because the number of observations in the Poisson regression is driven by both the number of patients and the dimension of the multivariate outcome for each patient, which is typically much larger than the number of procedure codes. Therefore, information matrices such as $\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)$ presented in Lemma 2.1 are not singular and can be inverted, and our estimator is not subject to the projection bias in high dimensions discussed in [15]. Our choice of ridge regression is purely for stabilizing estimates and not for solving the more common over-specification or singularity issue. Therefore, we can directly rescale the shrinkage estimator without the need to separately estimate a bias term.

2.3.3 Patient-level Clustering due to Simultaneous Analysis of Multivariate Outcomes

With the use of ridge regression we provide inference regarding the systematic associations with covariates such as healthcare system (site) for multiple CPT codes. Therefore, a vector of multivariate outcomes represents the data that is analyzed for each patient and characterizes the full set of CPT outcomes for that individual. Simultaneous regression with blocks of codes would then need to appropriately account for the multiple outcomes per patient. Standard GEE type sandwich variance estimates are a simple way to account for the potential within-subject correlation for this situation [16].

2.3.4 Provider-level Clustering

In Section 2.2.3 we discussed testing options when patients treated by the same provider are potentially correlated. In the same spirit, a sandwich variance estimator can be used to replace $\hat{\mathbf{\Omega}}_n$ discussed in Section 2.3.2 to generate a ridge test that accounts for provider-level clustering, which can be viewed as variations of a Generalized Estimating Equations (GEE) strategy [26]. Performance of the ridge test with and without a robust variance-covariance matrix is studied via simulation in Section 2.4.1.2.

2.4 Simulations

Previous research has compared group-wise association tests in the context of genome-wide association studies (see, for instance, [107, 17, 75, 9, 79]), and relevant results are detailed in Appendix A.7. Generalizing these results to our context implies that burden tests may have increased power to detect association at the block level when the direction of code-specific effects are similar. In contrast, when code-specific effects may differ in direction, SKAT has been shown to be an effective testing strategy [107]. When provider-level clustering is present, the L-KM method controls the type I error and increases power compared to competing methods [79].

We focus our simulation studies on characterizing the finite sample operating charac-

teristics of code-wise testing strategies where both common and rare codes are of interest. Sparse codes are likely to require exact methods to preserve the nominal type I error rate, while common codes are likely to require methods/models for overdispersed count data. We are not aware of any literature that provides a comprehensive characterization of test option performance in the CPT code context, and such evaluation is necessary to provide recommendations for routine comparison of healthcare utilization.

For all of our simulation studies, we consider four classes of testing options:

- Analysis of the CPT count data using a LRT and a ET based on either the simple Poisson model or the more general negative binomial model.

- Analysis of the derived any/none binary indicator using Binomial methods: Fisher's ET and Binomial LRT.

- Simple two-sample t-test with unequal variances, or with a sandwich variance estimator (z-statistic) for correlated data, as a potential semi-parametric method relying solely on moment assumptions.

- Ridge test that constructs p-value and confidence interval using debiased estimator assuming either independent or correlated data.

We evaluate the performance in terms of type I error rate and power across a full range of rate parameters as might be encountered in healthcare utilization data. We also evaluate inference of the ridge regression in terms of type I error and coverage.

We consider three types of underlying data for our simulation studies:

- $Y_{si}^c \sim$ Poisson(rate $= \lambda_s^c$) with $\mathrm{E}[Y_{si}^c] = \mathrm{Var}[Y_{si}^c] = \lambda_s^c$. Note that $\Pr(Y_{si}^c = 0) = e^{-\lambda_s^c}$.

- $Y_{si}^c \sim$ negative binomial($\mu = \lambda_s^c, r = 2$) with $\mathrm{E}[Y_{si}^c] = \lambda_s^c$ and $\mathrm{Var}[Y_{si}^c] = \lambda_s^c(1 + \frac{1}{2}\lambda_s^c)$. Note that $\Pr(Y_{si}^c = 0) = (\frac{r}{\lambda_s^c + r})^r = (\frac{r}{\lambda_s^c + 2})^2$.

- $Y_{si}^c \sim$ zero-inflated Poisson($\pi^c = \frac{r}{1+r} = \frac{2}{3}, \lambda_s^c$) with $\mathrm{E}[Y_{si}^c] = \lambda_s^c$ and $\mathrm{Var}[Y_{si}^c] = \lambda_s^c(1 + \frac{1}{2}\lambda_s^c)$. There are two components: $Y_{si}^c = 0$ with probability $\pi^c = \frac{2}{3}$, and $Y_{si}^c \sim$ Poisson(rate $= \frac{\lambda_s^c}{1-\pi^c}$) with probability $1 - \pi^c$. Both generate zero-counts, so $\Pr(Y_{si}^c = 0) = \pi^c + (1 - \pi^c)e^{-\frac{\lambda_s^c}{1-\pi^c}} = \frac{2}{3} + \frac{1}{3}e^{-3\lambda_s^c}$.

We dichotomize the individual count data to create indicators, Z_{si}^c, analyzed using a Binomial LRT. Note that $\Pr(Y_{si}^c = 0) \to 0$ as $\lambda_s \to \infty$ in the negative binomial and Poisson model implying that for count data with a large mean, the dichotomized data would be all ones and therefore be uninformative for evaluation of rate differences. Finally, for ETs we aggregate individual data, Y_{si}^c and Z_{si}^c, into cohort-level totals Y_s^c and Z_s^c.

Our choice of generating models is to allow both standard Poisson models as well as alternative distributions that generate overdispersed data. The primary role of the zero-inflated Poisson model is to allow evaluation of the flexibility or robustness of the negative binomial model for overdispersed data that are outside the assumed class of models used for analysis. We parameterize the three data generating models such that they all have the same mean λ_s, which represents the average number of times a procedure was delivered to a patient during a fixed (one unit) follow-up time period. In addition, the negative binomial model and the zero-inflated Poisson model are chosen to have the same variance.

In addition to concerns of sparsity, overdispersion, and model misspecification, we are also interested in the influence of provider-level clustering on validity of the tests, and the performance of testing strategies that account for correlated data. To this end we assign each patient to a provider randomly with an average cluster size of three patients per provider, and introduce correlation between patients within provider by including a provider-level random variable. We consider two types of mean-preserving random variables that fluctuate the mean λ_s by provider p:

- $\lambda_{s,p}^\beta = \lambda_s \cdot \gamma_p^\beta$, where $\gamma_p^\beta \sim$ Gamma(shape $= \frac{1}{\beta}$, scale $= \beta$) is shared by patients within the same provider p, with $\mathrm{E}[\gamma] = 1$, $\mathrm{Var}[\gamma] = \beta$, and $\mathrm{E}[\lambda_{s,p}^\beta] = \lambda_s$.

- $\lambda_{s,p}^\sigma = \exp[\log(\lambda_s) + b_p^\sigma]$, where $b_p^\sigma \sim \mathrm{Normal}(0, \sigma^2)$ is shared by patients within the

same provider p, with $\mathrm{E}[\lambda^{\beta}_{s,p}] = \lambda_s \cdot e^{\frac{\sigma^2}{2}}$ but $\mathrm{E}[\log(\lambda^{\beta}_{s,p})] = \log(\lambda_s)$.

2.4.1 Code-wise Test: Type I Error Rate

In the following sections, we assess size of the testing options under imbalance sample sizes, sparsity, overdispersion, model misspecification, and potentially correlated data introduced by provider behavior.

2.4.1.1 Independent data

In this section, we generate independent outcomes to estimate the type I error rate as the proportion of p-values less than the nominal α level of 0.05. We set $\log_{10}\lambda_0 = \log_{10}\lambda_1$ to range from -6 to 2, so that λ ($\equiv \lambda_0 = \lambda_1$) increases from 10^{-6} to 10^2 multiplicatively. Under this null the rate ratio is one and both cohorts have equal variances. For each scenario considered we conduct $5,000$ simulations with unequal samples sizes of $n_0 = 1,000$ and $n_1 = 3,000$ since this reflects the motivating example data.

The performance of each test varies by the average number of times a procedure was delivered in the sample, $(n_0+n_1)\lambda$, which we refer to as the frequency. We will discuss selected results presented in Figure 2.1 by four regions of $\log_{10}\lambda$. They are region I: $[-6,-4)$, region II: $[-4,-2)$, region III: $[-2,-0.5)$ and region IV: $[-0.5,2]$, which correspond to a frequency of less than 0.4, less than 40, between 40 and 1265, and over 1265, respectively, for the sample size of $4,000$. The comprehensive results with equal and unequal sample sizes are shown in Appendix A.8 Figures A.1 and Appendix A.2. We also evaluated coverage of the ridge test shown in Appendix A.8 Figure A.5.

For extremely low rates (region I) all methods have a type I error of nearly zero for any of the three types of simulated data. This is not surprising since a rare procedure that gets assigned to only one out of every $10,000$ patients should provide little information unless sample sizes are extremely large.

When outcome rates are rare (region II), with an expected rate λ of 0.1 to 10 per $1,000$,

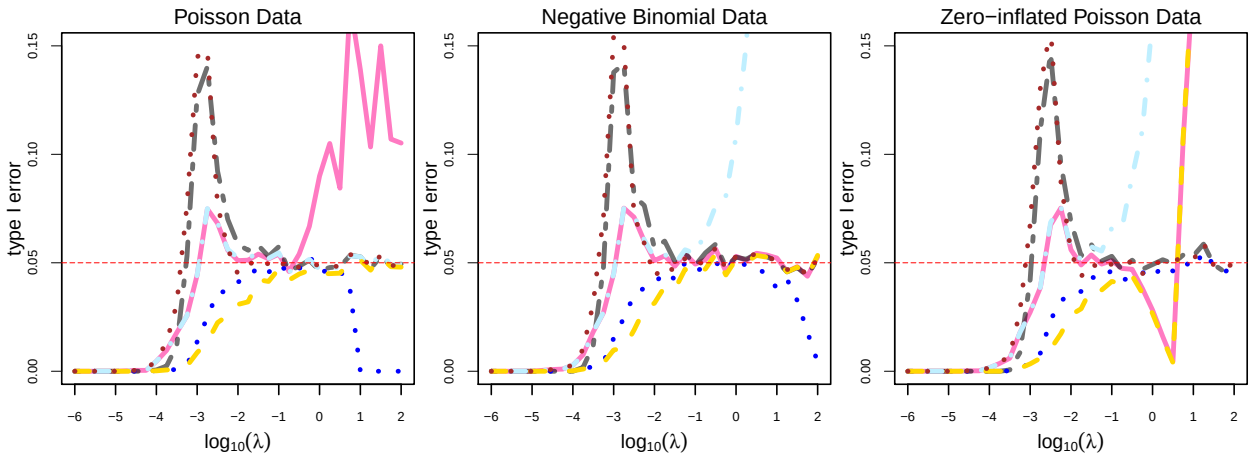

Figure 2.1: Type I error rates of CPT code-specific tests with unequal sample sizes ($n_0 = 1,000$, $n_1 = 3,000$) using Poisson data, negative binomial data, or zero-inflated Poisson data, each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale. Colored lines correspond to negative binomial LRT (——); negative binomial ET (– –); Poisson LRT (– – –); Fisher's ET (- - -); t-test (— —); ridge test (-·-·-).

or equivalently, a frequency of 0.4 to 40, we find that LRTs, t-test, and the ridge test have inflated type I error rates. Conversely, the three ETs still hold the type I error below the nominal level.

In region III when λ is approximately 100 per $1,000$, all of the tests perform well and have a type I error rate of around 0.05.

Finally, in region IV, the t-test and ridge test are the only two tests with a type I error rate near the nominal level for all three data-generating mechanisms. All of the model-based LRTs are subject to inflated type I error when the assumed model is incorrect. Specifically, the negative binomial LRT and ET break down when the true distribution is zero-inflated Poisson, although they are valid for both the distributions within their assumed class (negative binomial and Poisson). Also, when λ is large, the induced dichotomized data in the negative binomial and Poisson model are all ones and therefore the Binomial LRT and Fisher's ET will have a type I error rate of zero.

Our simulation results suggest that a valid and simple testing strategy for rate differ-

ences can be obtained from a dynamic test that uses the negative binomial ET if the total number of delivery in the sample (i.e., the frequency) is less than 40, and otherwise uses the semi-parametric t-test. In order to evaluate the performance of such a procedure we have calculated test size in additional simulations (in Appendix A.8 Figure A.6), and find that the dynamic test tracks the conservative type I error of exact methods for low rates, but then enjoys robustness to model assumption for moderate and large rates. In additional simulations we evaluated the threshold of 40 and find this appropriate with varying total sample size.

In summary, we find that no method can be reliably used across the entire spectrum of candidate rates that are encountered with CPT data. For rare rates exact testing methods are preferred, while for common rates robust methods such as the t-test and the ridge test perform well. Model-based count data LRT do not exhibit sufficient robustness to rare counts or model violation to be recommended for routine surveillance use.

2.4.1.2 Correlated data

In this section, we investigate influence of provider-level clustering, and evaluate performance of the proposed corrections in t-test, dynamic test, and ridge test that allow for correlation between patients within provider. Recall that we introduce correlation using two types of mean-preserving random variables (simulation settings in Section 2.4). Firstly, for a fixed λ, we study how strength of correlation influences the type I error by setting variances of the random variables, defined by β and σ^2, to increase from $\frac{1}{16}$ to 4 multiplicatively. Secondly, we fix β and σ^2, and set λ to range from 10^{-6} to 10^2 to study sensitivity and robustness to correlated data under both rare and common code scenarios.

Figure 2.2 shows selected results which are type I error rates of t-test, dynamic test, and ridge test using negative binomial data with provider-level clustering. For tests that ignores the correlation between patients within provider, we observe two patterns: when CPT codes are rare, the testing procedures are not very sensitive to correlation among patients, and the type I error is not substantially inflated; when CPT codes are common, type I error exceeds

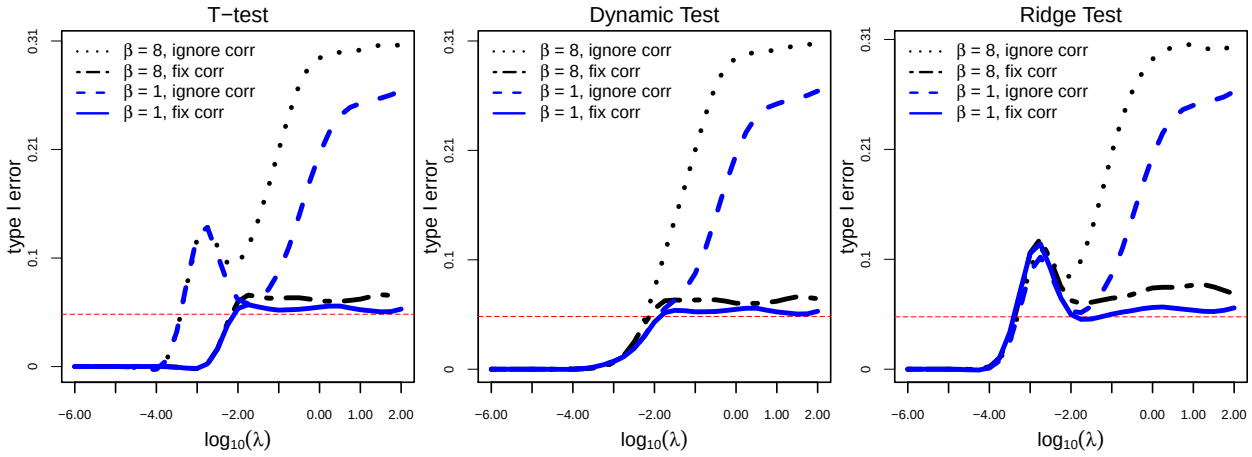

Figure 2.2: Type I error rates of t-test, dynamic test, and ridge test using negative binomial data with provider-level clustering. To introduce association, a mean-preserving random variable $\gamma_p^\beta \sim \text{Gamma}(\text{shape} = \frac{1}{\beta}, \text{scale} = \beta)$ with $\text{E}[\gamma_p^\beta] = 1$ and $\text{Var}[\gamma_p^\beta] = \beta$ is shared by patients treated by provider p. The cohorts have unequal sample sizes ($n_0 = 1,000$, $n_1 = 3,000$), each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale.

the nominal value. In addition, the type I error increases when either the utilization rate λ increases or the correlation controlled by β and σ^2 gets stronger. We also see that when the covariance matrix is estimated accounting for correlation within provider clusters in each of the tests, the type I error is corrected to the nominal α level. Comprehensive results for Gamma and Normal random variables, and across different data generating distributions including negative binomial, Poisson, and zero-inflated Poisson are quite similar and are shown in Appendix A.8 Figures A.3 and A.4.

2.4.2 Code-wise Test: Power

To explore power we focus on select event rates, $\lambda_0 = 1$ or 0.01, and empirically evaluate power as a function of various rate ratios, $\lambda_1 = \text{RR} \cdot \lambda_0$. We let $\log_2 \text{RR}$ range from 0 to 3, so that the RR increases from 1 to 8 multiplicatively, and we conduct $2,000$ simulation replications for each situation. Figure 2.3 shows selected results for unequal sample sizes similar to our motivating data, and show a monotone increase in power with increasing RR,

and a small loss of power for exact and robust methods relative to the correctly specified LRTs. Comprehensive results for equal and unequal sample sizes are quite similar and are shown in Appendix A.8 Figures A.7 and A.8. A few observations are notable, with the first being that application of Fisher's ET when the event rates are large may result in a substantial loss of power. When $\lambda_0 = 1$ we see that most test procedures achieve power greater than 80% for $\log_2 \mathrm{RR} > 0.25$ for all three data-generating mechanisms, while Fisher's ET has power $< 50\%$ for all RR values under a zero-inflated Poisson model. However, with lower event rates (e.g., $\lambda_0 = 0.01$) we see a small reduction in power with use of ETs. Therefore, the power plots reinforce recommendations based on preservation of test size: ETs appear valid and reasonably powered for low event rates; while robust methods are valid and retain power for common event rates.

2.5 Application: Comparing Healthcare Utilization between Henry Ford Health System and Kaiser Permanente

The development and evaluation of statistical methods for comparing rates of medical procedure utilization is motivated by the Back pain Outcomes using Longitudinal Data (BOLD) project, which enrolled 5239 patients aged 65 years and older with a new episode of back pain [50]. In order to enroll more than $5,000$ patients, recruitment was conducted from three healthcare systems. Primary scientific questions focus on medical interventions such as early radiologic imaging and subsequent patient reported pain and function outcomes. In order to combine EMR data across the three systems we need to understand any differences in procedure endorsement across the sites. Therefore, we focus on CPT coding data across all domains including imaging, laboratory, and diagnostic procedures. Specific sub-cohorts can be defined using the demographic or clinical information. Here we focus on the cohorts defined by the enrollment site for a patient since both geographic and healthcare system differences may be associated with different CPT coding patterns. We compare healthcare utilization between two largest sites: Henry Ford Health System in Detroit and Kaiser Permanente in Northern California, which include 4040 patients.

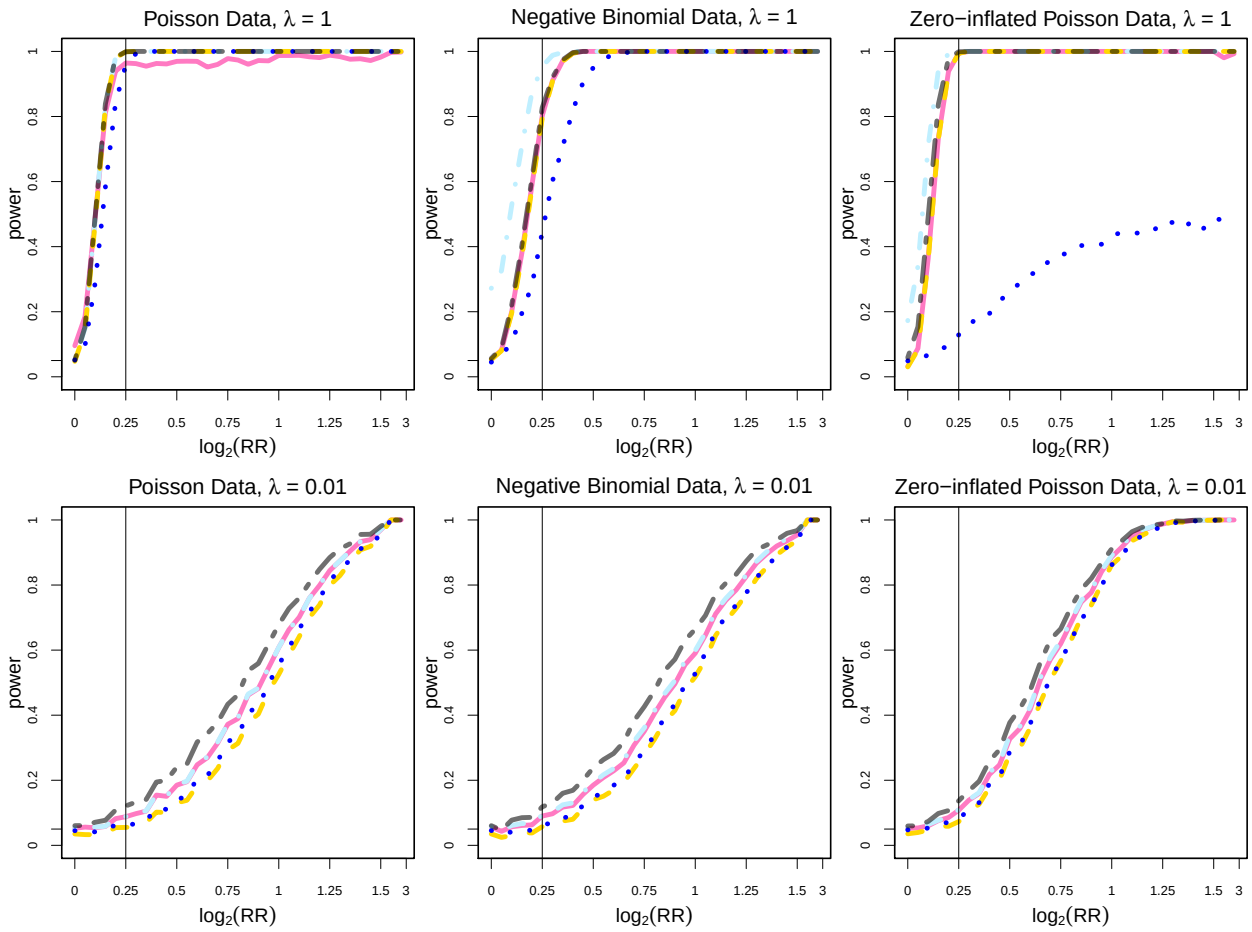

Figure 2.3: Power of CPT code-specific tests with unequal sample sizes ($n_0 = 1,000$, $n_1 = 3,000$) using negative binomial data, Poisson data, or zero-inflated Poisson data, each with a mean of $\lambda_0 = 1$ or 0.01 in cohort 0, and a rate ratio on $\log 2$ scale ($\log_2 \frac{\lambda_1}{\lambda_0}$) ranging from 0 to 1.5. Colored lines correspond to negative binomial LRT (——); negative binomial ET (— —); Poisson LRT (– – –); Fisher's ET (- - -); T-test (— – —).

2.5.1 Hypothesis Testing

First, we investigate the significance of the difference in code utilization for individual codes and blocks of codes defined by the CCS-Services and Procedures. We use the dynamic test for both code-specific comparison and the burden test; for block-based inference we use both the burden test and SKAT. We present code-wise p-values on the $-\log_{10}$ scale for all CPT codes in a "Manhattan plot" (Figure 2.4), for which the codes in a block are contiguous and plotted with the same color. We truncate any p-value at 10^{-17}. We also add the group-wise p-values to the Manhattan plot, one for each block. We include two horizontal lines which are the Bonferroni corrected significance thresholds for code-level and block-level comparisons. There are a total of $C = 2,424$ CPT codes with non-zero counts, and based on the CCS classification we have $B = 192$ blocks.

Figure 2.4(a) shows that there are many codes with utilization differences that are statistically significant, and that these codes tend to cluster in select domains. We detected significant difference among 31 out of 192 blocks using the burden test, and only 5 using the SKAT. Specifically, in 27 blocks, the burden test rejects the null hypothesis while SKAT does not, and there is only 1 block for which the SKAT has a significant result in contrast to the result from the burden test.

We zoom in on select blocks to investigate detailed patterns as shown in Figure 2.4(b). The burden test for "Laboratory - Chemistry and Hematology" rejected the null hypothesis, while the SKAT p-value did not reach the block-wise significance level. We find that this is driven by the abundance of codes whose utilizations are consistently higher at one site, a situation in which the burden test has higher power.

In the block called "Mammography", although both the burden test and SKAT give a significant result, SKAT gives a much smaller p-value than the burden test. Looking into the data, we see that Henry Ford Health System uses exclusively so-called G codes for recording of mammography, while Kaiser Permanente uses the common numeric codes for mammography. We learned that G-codes refer to digital mammography for screening

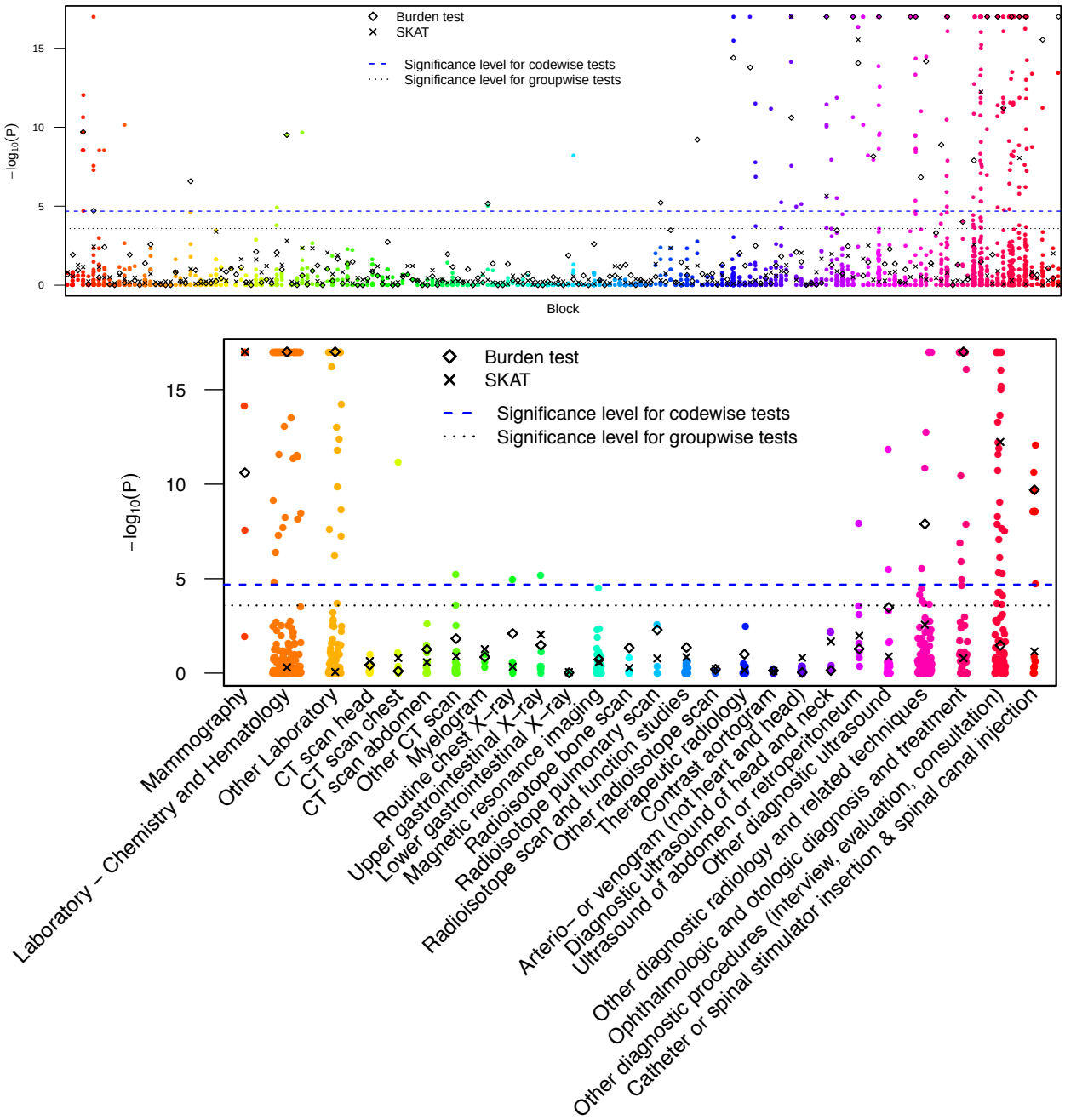

Figure 2.4: A full Manhattan plot for code-wise comparison of Henry Ford Health System and Kaiser Permanente plotted by block, overlaid with results from the group-wise comparison using the Burden test and SKAT for each block. The y-axis is truncated at a p-value of 10^{-17}. Bonferroni corrected significance levels for code-wise and group-wise tests are shown. Panel (b) is a zoom in version of the Manhattan plot for select blocks.

or diagnosis, whereas numeric codes refer to non-digital (film) mammography. Therefore, if the cost of breast cancer screening was the focus of analysis, then minor differences in codes may be important to capture. On the other hand, if overall interest is in the rate of patient screening regardless of imaging technology, then G-codes and numeric codes may be combined to define a general mammography procedure and the detailed differences will not be important. Such distinctions relate to the choice of code-specific or block-specific inference procedures that we present, as well as the trade-off between sensitivity and power as illustrated by comparing the results from the burden test and from SKAT.

Our methods are intended to identify specific codes, or groups of codes, that appear to have different recorded utilization. Additional investigation is required to separate whether the finding corresponds to actual differences in patient care, or whether coding variation through use of alternative codes may explain differences in observed endorsement rates. In our use of these methods with our healthcare delivery system studies, we have used the signals from our testing procedure to engage in discussion with the individual systems to ultimately attribute observed differences to practice or coding variation.

2.5.2 Rate Ratio Estimation and Inference

A key aspect of understanding differences in CPT endorsement is the direction and magnitude of rate differences. Therefore, we also estimate rate ratios for all codes simultaneously to compare the utilization pattern in Henry Ford Health System to Kaiser Permanente, adjusting for age category, sex, and race. Due to potential sparse codes we use penalized Poisson regression as detailed in Section 2.3, and we provide inference for ridge regression using methods detailed in Section 2.3.2. In the BOLD study, patients were recruited through primary care clinics, and information on their primary care provider is available. To illustrate the practical use of our proposed methods that account for provider-level clustering discussed in Section 2.3.4, we adopt the primary care provider as our level of clustering which is appropriate for the study design. In BOLD study we have 4040 patients and 1819 providers.

To display the point estimation results we use dynamic graphical methods that plot the

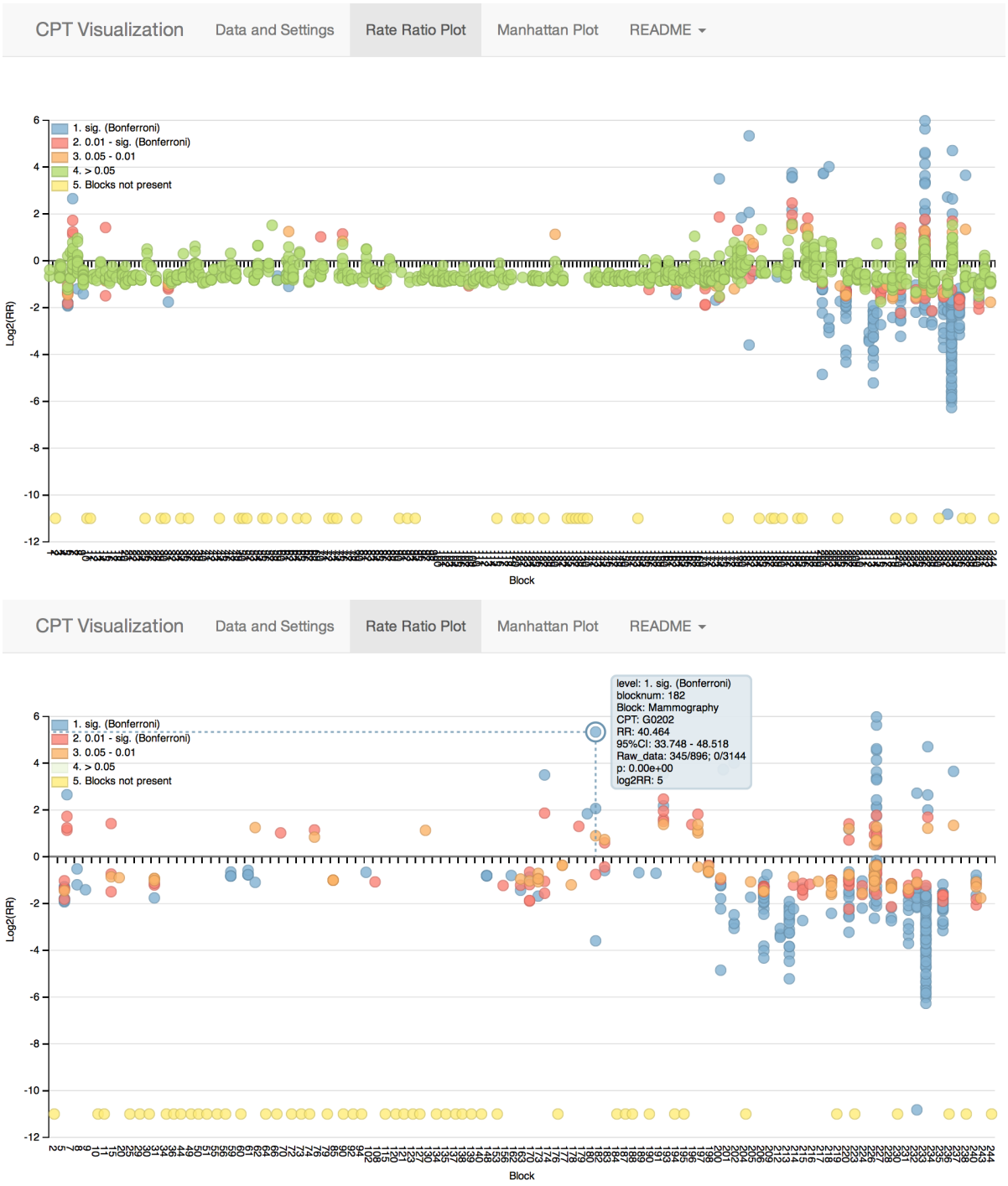

Figure 2.5: Rate ratio estimates comparing Henry Ford and Kaiser Permanente for each CPT code based on a penalized Poisson regression with ridge penalty. Code-specific rate ratios ($\log_2$ scale) are plotted against the block that the each code belongs to, color-coded according to four levels of p-values: $(0, \alpha]$, $(\alpha, 0.01]$, $(0.01, 0.05]$, and $(0.05, 1]$, where α is the Bonferroni corrected significance level. The plot function can dynamically provide additional information for each point showing the block, the code, the rate ratio, the 95% confidence interval, the p-value, and the raw data, as illustrated with one point in panel (b).

estimated code-specific rate ratios on a $\log_2$ scale versus the block to which each code belongs to (Figure 2.5). We color-code each point according to the significance of the code from the hypothesis test using the ridge test detailed in Section 2.3.2. The p-values are split into four regions, each corresponding to a color. They are: $(0, \alpha]$, $(\alpha, 0.01]$, $(0.01, 0.05]$, and $(0.05, 1]$, where α is the Bonferroni corrected significance level. The summary plot displays the healthcare utilization pattern over the entire spectrum of all possible procedures, but detail on individual codes can be revealed using dynamic graphical methods, and as shown in Figure 2.5, we display both statistical information and medical details for any given CPT code. For blocks that are not present in the data, we plot place holders at the bottom using a unique color. We note that individual CPT code rate ratios are shrunk to a block level rate ratio due to our choice of penalization that incorporates the block-code hierarchical structure.

There are a variety of factors that may drive measured utilization differences, including healthcare practices, coding regulations, data quality issues, and differences in patient characteristics. Our methods can reveal differences that should be followed up with additional investigation into the underlying drivers for the observed differences. To this exploratory end, we make the plot interactive which displays the tooltip detailing the information of the block, the code, the rate ratio, the 95% confidence interval, the p-value, as well as the raw data. One can also filter on select results based on p-value categories. Pointing to a specific point in the plot, for example, the G0202 in the "Mammography" block, we can see from the raw data that Henry Ford uses the G codes exclusively, while Kaiser Permanente uses the five-digits codes, as is discussed in Section 2.5.1. Such pattern drives the estimated rate ratio to be high and the p-value to be low.

The summary rate ratio plot serves as an interpretable tool for clinical researchers and data managers to explore healthcare utilization patterns among sub-cohorts. We have used this tool as part of our data quality control, and for providing potential alternative explanations that need to be considered in comparative utilization analyses across observed patient subgroups. We implement the interactive plot in both an R package and a shiny appli-

cation. The shiny app is available online at `https://xu-rita-shi.shinyapps.io/CPT_visualization/`.

2.6 Discussion

Contemporary biomedical research is now leveraging the electronic medical records for both comparison of alternative treatment options and to generate individual predictions. Increasingly there are large networks of hospitals or healthcare systems that are assembled to provide sizable cohorts. With these efforts comes the need to compare patterns of utilization across sub-groups within modern cohorts, either to understand systematic issues with respect to data quality or coding variation, or to compare utilization across patient subgroups defined by treatment or medical indication. Therefore, we have developed multilevel hypothesis testing and rate ratio estimation methods that can be used either for evaluation of potential data issues or for comparative inference.

First, we detailed statistical testing methods for evaluating differences in procedure assignments between two groups, and provide inference at both the code and block levels. To compare utilization at the code level, we discussed potential likelihood ratio test and conditional exact tests. We focused on three candidate distributions: the Poisson, negative binomial, and Binomial distribution if data are dichotomized. When comparing rare procedure codes which might lead to low power and violation of assumptions for the asymptotic χ^2 approximation, the conditional exact test provides a viable option. We also considered semi-parametric testing using the two-sample t-test. We learned from our simulation study that different tests work well in different scenarios, and the dynamic test tracks the conservative type I error of exact methods for low rates, but then enjoys robustness to model assumption for moderate and large rates. To compare utilization on block level, we transferred methods from genome-wide association studies to the EMR context, including the burden test and the Sequence Kernel Association Test. Both the burden test and SKAT evaluate utilization patterns by combining a block of similar codes, which may substantially increase power for rare codes in particular.

Second, we detail estimation and inference of utilization rate ratios via penalized Poisson regression with a tailored form of penalty that takes advantage of the hierarchical structure of the CPT codes. Our proposed method shrinks the code-specific estimates to the block level, effectively borrowing information from all other codes within the same block. Such shrinkage is especially important for rare codes for which individual rate ratio estimates may be highly unstable. We also develop inference methods that account for shrinkage bias and construct statistical tests (p-values) and confidence intervals using the distribution of a debiased estimate.

Third, we consider provider behavior as an important driver of patterns in healthcare utilization and expanded the inference method to account for potential correlation within provider. We learned from simulation that for rare CPT codes, testing is not sensitive to provider clustering, because there is not much information to be influenced by within-provider correlation. In contrast, correlation does inflate the type I error among common outcomes, and the amount of inflation increases with the strength of the correlation and the mean of the outcome. We are also able to control the inflated type I error under correlated data back to its nominal value by correcting the variance-covariance estimate.

Finally, we ultimately provide interpretable dynamic graphical tools that can help researchers to explore and interpret the healthcare utilization patterns. We use a CPT code version of the genomic Manhattan plot to display testing results, and we use an interactive plot to present both the significance and the magnitude of rate differences. The interactive plot enables us to see useful global information and select detailed information that facilitates discovery of key utilization differences. Although we focus on CPT coding differences, the general testing and estimation framework is also applicable to other forms of structured EMR data such as diagnostic coding data (ICD-9, or ICD-10) and is particularly useful when any coding system can be hierarchically organized.

A potential limitation of our work is the need to further investigate whether any statistical finding corresponds to actual differences in patient care, or whether coding variation through use of alternative codes may explain differences in observed endorsement rates. Another

limitation is that in our data application, we only had primary care provider IDs and did not have detailed specialty care provider information to illustrate how one could use an extension of the GEE type variance estimator to accommodate non-nested clustering. In addition, although we consider accounting for patient-specific follow-up time to partially account for missing data, tailored methods targeting missing data is an important future direction in the use of EMR data. Lastly, the longitudinal nature of EHR data has great potential for research to understand the temporal changes in patient treatment history and patient health status, which requires future work.

Chapter 3

EVALUATION AND EXTENSION OF LITERATURE ON REGRESSION ADJUSTMENT OF THE PROPENSITY SCORE

In this chapter, we provide a brief review on methodology and theory of regression adjustment using the propensity score. We start with introducing the key concepts in causal inference and the propensity score. Then we detail literature on both the methodology and theory of regression adjustment using the propensity score as a covariate.

3.1 Causal Inference and the Potential Outcomes Framework

Causation is inferred by any difference between the outcomes under exposure and control, when all other circumstances are the same. Accordingly, for each subject i in the target population, we define a pair of variables $\big(Y_i(1),\ Y_i(0)\big)$ as the outcomes that would be observed under exposure and control, respectively. Denote the binary exposure as T_i, $i = 1, \ldots, n$ for subject i, taking on value 1 (exposed) or 0 (unexposed). For each subject, only one of the potential outcomes is observed, i.e., the observed outcome $Y_i = Y_i(1)$ if exposed $(T_i = 1)$, and $Y_i = Y_i(0)$ if unexposed $(T_i = 0)$, with $Y_i(1) \mid T_i = 0$ and $Y_i(0) \mid T_i = 1$ missing.

The gold standard for estimating a causal effect is to conduct a fully randomized controlled experiment, in which the exposed and unexposed groups are balanced. In this case, the mean of observed outcomes in the exposed group, $E[Y \mid T = 1]$, will be equal to the mean of potential outcomes in the entire population (both the exposed and unexposed groups), $E[Y(1)]$. Thus, one can directly estimate the population average using the observed portion.

In observational studies, however, differences in the outcomes between the two arms could be due to both pre-existing systematic differences and the drug effect. In the presence of confounding effects in observational studies, [87] proposed the strongly ignorable exposure

assignment assumption: $\big(Y(1),\ Y(0)\big) \perp T \mid X$, where X denotes the baseline covariates. In other words, we assume that there is no unmeasured confounder, and thus treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates. This assumption allows one to estimate the conditional mean of potential outcomes $Y(1)$ using the observed portion $Y \mid T = 1$, as if one conducted randomization within each stratum of X, i.e., $E[Y(1) \mid X] = E[Y \mid T = 1, X]$.

Causal inference is a comparison of the population-level or marginal averages of the potential outcomes. The most common form of comparison is the mean difference, i.e., the causal exposure effect is measured by the average treatment effect (ATE), $ATE = E[Y(1)] - E[Y(0)]$, or the average treatment effect on the treated, $ATT = E[Y(1) \mid T = 1] - E[Y(0) \mid T = 1]$. With estimating the population average of potential outcomes as the ultimate goal, causal inference methods either provide a balanced population that mimics one from a randomized experiment, or impute the unobserved potential outcomes. A selected review of causal inference methods will be provided in Section 4.3.

3.1.1 The Propensity Score

The propensity score is the probability of being exposed or treated given the subject's characteristics, i.e., $S \equiv h(X) = E(T = 1 \mid X) = P(T = 1 \mid X)$ [87]. It has two important roles. First, it is a summary score that reduces the dimension of the baseline covariates: it summarizes a vector of the baseline covariates into a one-dimensional variable according to how they describe exposure-proneness. Second, it is a balancing score: conditional on the propensity score, the baseline covariates are similar between exposure and control groups, which ensures that $E[Y \mid T = 1, S] = E[Y(1) \mid S]$. An advantage of the propensity score is that one can fit a non parsimonious and over-parameterized propensity score model. However, definition of the goodness of a propensity score was inconsistent in the literature. [1] defined the criterion as to achieve the best estimated probability of treatment assignment, i.e., goodness of fit. On the other hand, [94] discussed that the propensity score model should not be evaluated based upon goodness of fit or its discrimination, but whether it adequately

balances the confounders. Although the latter is more commonly seen in literature, we observe that for methods that have good asymptotic properties when the estimated propensity score is a consistent estimator of $E(T = 1 \mid X)$, such as the IPTW and the AIPTW estimator, it is preferred to have the best estimated probability of treatment assignment that provides unbiased prediction of the mean of treatment [90]. In contrast, matching and stratification only requires balance in covariates conditional on the propensity score, in which case the propensity score model can be evaluated by how well it balances the covariates in the treatment and control groups, instead of how well it predicts the treatment assignment.

The following sections in this chapter covers literature review on both the methodology and theory of regression adjustment using the propensity score as a covariate. We also provide asymptotic analyses for one particular estimator in the class of estimators that utilizes regression on the propensity score as a covariate.

3.2 Literature Review on Methods using Regression on the Propensity Score

3.2.1 Introduction

Regression adjustment for propensity score has been widely used in observational studies as a technique to balance covariates, an analog of randomization in experiments [90]. The rationale behind propensity score adjustment is the balancing property: conditional on the propensity score S, the distribution of the covariates X is independent of the exposure T [87]. If one further assumes that there are no unmeasured confounders, then substituting X with a one-dimensional scalar S, is sufficient to provide a consistent estimate of the marginal means of the potential outcomes. This is because given the propensity score S, the covariates X act as a precision variable and thus sufficient control of confounding is achieved by adjustment of the propensity score only [42]. In other words, the mean potential outcomes $E[Y(1)]$ (and $E[Y(0)]$) can be identified by $E[E(Y \mid T = 1, S)]$ (and $E[E(Y \mid T = 0, S)]$).

In the following subsections, we provide a review of literature on estimation of causal effects via regression adjustment of the propensity score estimated in a preliminary step. We

first consider adjusting for the propensity score as a linear term in the outcome regression model, and discuss conditions for unbiasedness under misspecification of the propensity score model or the outcome regression model, and the distinction between collapsible and non-collapsible models. We then discuss the literature on nonlinear adjustment of the propensity score, under the assumption that the propensity score is correctly specified and thus has the balancing property.

3.2.2 Linear Adjustment

In this section, we consider the properties of treatment effect estimators when one regresses the outcome on the treatment and the propensity score as a covariate, i.e., linear adjustment using the propensity score as a covariate. We consider the scenarios of collapsibility, whether the propensity score is known or correctly specified, and whether the outcome is truly determined by a simple linear term of the propensity score. We summarize the consistency of treatment effects in Table 3.1, which will be detailed in the following sections.

Table 3.1: Consistency of Treatment Effects Estimated via Linear Adjustment of the Propensity Score. "PS✓" denote using known or correctly specified propensity score; "Y|T,PS✗" denote that linear adjustment of the propensity score is a misspecification of the outcome model.

Outcome Model		PS✓ Y|T,PS✗	PS✗ Y|T,PS✗
Collapsible (log-) linear	no interaction	✓	✓ (if T⊥ X|PS)
	interaction	✗	✗
Noncollapsible	conditional trt eff	✗ (attenuated) X: precision var	✗
	marginal trt eff	✗ (poor prediction of E[Y|T,PS])	

3.2.2.1 Linear (or Log-Linear) Outcome Regression

Unlike odds ratios (ORs) or hazard ratios (HRs), the risk differences (RD) and rate ratios (RR) are collapsible, if there is no effect modification [34, 33]. That is, in a linear (or log-linear) regression without effect modification, conditional treatment effect is equal to the marginal effect. In such a situation, including the treatment variable and the *known* propensity score (or an estimated propensity score from a *correctly specified* model) in the model has been shown to yield unbiased estimate of the treatment effect, even if propensity score should not appear as a linear term in the model, i.e., the model $E[Y|T,S]$ is mis-specified. [37] and [101] provided two versions of proofs. An intuitive explanation is that because the treatment is independent of the covariates conditional on the true propensity score [87], if one includes extra covariates in the model, the additional covariates serve as precision variable that are predictive of the outcome but not correlated with the treatment variable. Therefore, even if one mis-specify $E[Y|T,S]$, the residual only influences the efficiency but not the consistency.

However, when the propensity score model is *unknown or mis-specified*, adjusting for the estimated propensity score as a linear term has been shown to lead to some bias. In particular, [37] provided a closed form of the bias term

$$\sqrt{\frac{\sum_{i=1}^{n}[\alpha(X_i)-\overline{\alpha(X)}]}{\bar{T}(1-\bar{T})}}\cdot\frac{\rho_{tx}-\rho_{ts}\rho_{xs}}{\sqrt{(1-\rho_{ts}^2)(1-\rho_{xs}^2)}}\cdot\sqrt{\frac{1-\rho_{xs}^2}{1-\rho_{ts}^2}}\cdot\beta_x, \tag{3.1}$$

where $\alpha(X)=E[Y\mid T=0,X]=\beta_0 T+X\beta_x$, β_x is the coefficients of X in $\alpha(X)$, ρ_{tx} is the correlation between treatment T and $\alpha(X)$, ρ_{ts} is the correlation between treatment T and the propensity score S (on the scale that is used in the model), and ρ_{xs} is the correlation between $\alpha(X)$ and S.

The above bias term is zero when one of the following three conditions hold:

(1) the covariates are not prognostic given the treatment;

(2) the treatment is independent of the covariates;

(3) the prognostic score $\alpha(X)$ and the propensity score $S = \gamma(X)$ have perfect linear correlation, and $cov(T, \alpha(X)) = cov(T, \gamma(X))$.

Note that it is sufficient for condition (3) to hold when the outcome model adjusting for treatment and linear term of the propensity score is correctly specified.

There is in fact one additional condition, which leads to unbiasedness even if the propensity score is mis-specified and the outcome depend on the propensity score in a nonlinear fashion. Based on the proof in [37], and noting that $\frac{\rho_{tx}-\rho_{ts}\rho_{xs}}{\sqrt{(1-\rho_{ts}^2)(1-\rho_{xs}^2)}}$ is zero when $T \perp X \mid S$ (and thus $T \perp \alpha(X) \mid S$), we see that as long as the estimated propensity score balances the covariates perfectly, i.e., the partial correlation between treatment and the covariates given the propensity score is zero, then linear adjustment of propensity score in a linear or log-linear model without effect modification yields unbiased estimate.

It is important to note that when there is effect modification, i.e., $E[Y(1) \mid S] - E[Y(0) \mid S] \neq$ constant, the above unbiasedness does not hold.

3.2.2.2 Nonlinear and non-collapsible Outcome Regression

When one is interested in a conditional treatment effect, linear adjustment of the true or estimated propensity score in a nonlinear model may lead to bias, seen in simulation study of [4] looking at estimates of conditional odds ratio and hazard ratio.

Intuitively, if the relationship between the outcome and the propensity score is mis-specified, we can think of the correction term $(\alpha(X) - \beta_S \cdot S)$ as an omitted function of confounders. According to [87], conditioning on the true propensity score will make the confounders behave like "precision variables" – ones that are independent of the treatment variable and associated with the outcome. In a non-collapsible model, failure to adjust for a "precision variable" will attenuate the estimates towards the null. The amount of the attenuation depends on both the magnitude of the adjusted coefficient of treatment and the magnitude of the coefficient for the precision variable [34]. A detailed discussion is available in the study of [103], which decomposed the outcome regression model into the

propensity score and a remainder term, and studied the effect of omitting the remainder term in collapsible and non-collapsible models.

When one is interested in a marginal effect that has a causal interpretation, regression adjustment for propensity score as a linear term followed by standardization can lead to biased estimate. This class of methods uses regression adjustment for the propensity score to predict the missing potential outcome in order to estimate the mean of potential outcomes. To obtain a good prediction, the relationship between the outcome and the propensity score may not be linear. Linear adjustment of the propensity score will thus lead to poorly predicted potential outcome and thus biased mean potential outcomes.

[5] and [7] looked at the bias of linear adjustment of propensity score for estimating the marginal odds ratio and marginal hazard ratio. However, the author took the adjusted coefficient of the treatment as the estimate, which is in fact the conditional effect rather than the marginal effect. The conditional effect has an causal interpretation only when the model is collapsible, which is not the case for Cox PH model and logistic regression model. There is need for extensive investigation of performance of the class of estimators using linear and nonlinear adjustment of propensity score in a nonlinear and noncollapsible model looking at bias of the estimated marginal treatment effect.

3.2.3 Nonlinear Adjustment

In many cases, the propensity score and the outcome may not have a linear relationship. Including the propensity score as a linear term may therefore be inappropriate. To address this issue, efforts have been made to relax model assumptions and allow a nonlinear relationship between the outcome and the propensity score. We now review the literature on nonlinear regression adjustment of the propensity score method, with the assumption that the propensity score is correctly specified and thus is sufficient for control of confounding.

In the context of missing data, [56] proposed to adjust for the propensity score nonparametrically using a penalized spline model, and in addition adjust for the residuals from projecting each of the covariates on the propensity score, called the "Propensity Penalized

Spline Prediction (PSPP)" method. The spline function provides flexibility in adjustment of the propensity score, and the parametric adjustment of the residuals gains efficiency and control for minor residual confounding. However, the paper focused on the limited scenario of a continuous outcome and the parameter of interest is the marginal mean of the outcomes. Further investigation on deployment of the PSPP method in causal comparison with broader outcome distributional families is needed.

[36] proposed the "multiple imputation (MI) with two subclassification splines (MITSS)" method, which is a Bayesian version of the PSPP approach applied to the causal inference context, based on estimating two posterior response surfaces of the pair of potential outcomes. The authors studied methods for both binary and continuous outcomes. They proposed estimation of both causal effect and standard error using MI. They showed via simulation study that the MITSS method is the most efficient and has coverage levels that are closest to nominal level, compared to matching, subclassifications, weighting and covariate adjustment in regression.

In the spirit of flexible adjustment of propensity score, [93] suggested to classify the propensity score into quantiles and include in the regression as additional dummy variables, which is equivalent to fitting a nonparametric step function of the propensity score as an additional variable. They also discussed the fact that in a generalized linear model, the regression coefficient no longer corresponds to the regression estimate of the mean difference between outcomes. They have also mentioned potential covariate-by-treatment interaction and estimation of robust standard errors. Simulation compared performance of methods showed that including summaries of propensity score as additional covariates may improve the performance. However, their work is limited to linear regression for a continuous outcome, simulation on non-collapsible models is needed to provide a comprehensive comparison.

The aforementioned three methods include a nonlinear function of the propensity score as an additional covariate in addition to the baseline covariates, which can be hard to fit when the outcome is rare. In contrast, [68] considered including the treatment and a nonlinear function of the propensity score as the only two covariates additively in a generalized additive

model (GAM). The nonlinear function of the propensity score was approximated as a linear combination of thin plate regression splines. Simulation has shown that such approach outperforms stratification in terms of both bias and efficiency. However, like the one in [56], their development was restricted to collapsible models without effect modification, in which the average treatment effect is equal to the coefficient of the treatment variable, i.e., there is no need to further estimate the pair of mean potential outcomes.

A further generalization of GAM method is considered in [29], which compared propensity-based estimators of the marginal relative risk in healthcare database studies with rare outcomes. Estimating two outcome spline functions of the propensity score was considered, which shares several commonalities with the MITSS method although no additional covariate was adjusted. The simulation results showed that the spline regression on the propensity score method provide lower bias and mean squared error in the context of rare binary outcomes, regardless of the propensity score model estimation method. However, there is no formal statistical representation and/or analyses of theoretical properties of such an estimator.

3.3 Theoretical Properties of Regression on the Propensity Score

Studying the $\sqrt{n}$ consistency, asymptotic normality, and semiparametric efficiency can provide not only comprehensive understanding of the relative performance of the estimators, but also valid statistical inference and guidance on the choice of smoothing parameters when we use nonparametric regression on the propensity score. Among the previous papers on methodologies using the idea of regression adjustment for propensity score, the asymptotic properties of the estimators were barely studied, except that [56] has shown consistency of the PSPP estimator.

In this section, we first discuss the role of the propensity score in identification of the parameter of interest and the efficiency in estimation. We then briefly review the critical work of [72] and introduce pathwise derivative and semiparametric efficiency theory, which has been widely used in statistics and econometrics.

We will discuss one particular application of the semiparametric efficiency theory in the estimation of the ATE. We start with derivation of the efficient influence function of the G-computation parameter [39]. Then we discuss the influence function representation of another identification of the ATE, which is the nonparametric regression on the nonparametrically estimated propensity score followed by standardization [41]. We finally discuss the regularity conditions under which the three-step estimator is asymptotically normal with a variance achieving the efficiency bound, which have been detailed in [60] for kernel regression. For spline regression, we provide an extension of the discussion in [40] and characterize required regularity conditions.

3.3.1 The Dual Identification of the Mean of Potential Outcomes

Because the propensity score is a one-dimensional balancing score, a key idea is to substitute the propensity score for the set of covariates in an outcome regression model, which allows both control of confounding and dimension reduction. In other words, the mean of potential outcomes $E[Y(\cdot)]$ (or the ATE) can be identified in two ways.

Traditionally, the mean of potential outcome is identified by

$$E[Y(t)] = E\Big\{E[Y \mid T = t, X]\Big\},$$

and

$$\text{ATE} = E\Big\{E[Y \mid T = 1, X]\Big\} - E\Big\{E[Y \mid T = 0, X]\Big\},$$

which has also been referred to as the G-computation parameter.

Alternatively, the mean of potential outcome can also be identified by

$$E[Y(t)] = E\Big\{E[Y \mid T = t, S]\Big\}.$$

and

$$\text{ATE} = E\Big\{E[Y \mid T = 1, S]\Big\} - E\Big\{E[Y \mid T = 0, S]\Big\},$$

which motivates the method of regression adjustment using the propensity score followed by standardization.

A relevant property that is frequently used in theoretical analyses is, (taking $E[Y(1)]$ as an example)

$$E[Y(1)] = E\left[\frac{TY}{S}\right] = \begin{cases} E_X\{E[\frac{TY}{S} \mid X]\} = E_X\{E[\frac{T\cdot E[Y|T=1,X]}{E[P(X)|X]}]\} = E_X\{E[\frac{T\cdot E[Y|T=1,X]}{S}]\} \\ E_X\{E[\frac{TY}{S} \mid P(X)] = E_X\{E[\frac{T\cdot E[Y|T=1,S]}{E[S|P(X)]}] = E_X\{E[\frac{T\cdot E[Y|T=1,S]}{S}]\}. \end{cases}$$

Similarly, we have

$$E[Y(1)] = E\left[\frac{TY}{S}\right]$$

$$= \begin{cases} E_X\{E[\frac{TY}{S} \mid X]\} = E_X\{E[\frac{E[T|X]\cdot E[Y|T=1,X]}{E[P(X)|X]}]\} = E_X\{E[\frac{S\cdot E[Y|T=1,X]}{S}]\} = E_X\{E[Y \mid T=1,X]\} \\ E_X\{E[\frac{TY}{S} \mid P(X)] = E_X\{E[\frac{E[T|X]\cdot E[Y|T=1,S]}{E[S|P(X)]}] = E_X\{E[\frac{S\cdot E[Y|T=1,S]}{S}]\} = E_X\{E[Y \mid T=1,S]\}, \end{cases}$$

which in fact indicates the dual identification:

$$E[Y(1)] = E\left[\frac{TY}{S}\right] = \begin{cases} E_X\{E[Y \mid T=1,X]\} \\ E_X\{E[Y \mid T=1,S]\}. \end{cases}$$

3.3.2 The Value of Knowledge about the Propensity Score in Efficiency

Note that the ATE $= E[Y(1)] - E[Y(0)]$ does not depend on knowledge about the distribution of $T \mid X$, i.e., knowledge about the propensity score is nuisance. Here we summarize a few facts that are critical to propensity score methods as follows:

1. Knowledge of the propensity score does not change the efficiency bound for estimating average treatment effects [39]. For already-efficient estimators, using known or estimated propensity score thus does not change the efficiency of the estimator. For inefficient estimators such as the IPTW, using the estimated propensity score does improve the efficiency.

2. Despite the powerful dimension reduction property, it is well known that adjusting for the *known* propensity score will lead to an inefficient estimator of the average treatment effect with lower asymptotic efficiency than one based on adjusting for all covariates nonparametrically.

3. Interestingly, as discussed in [41], although nonparametric regression on the known propensity score leads to an inefficient estimator, nonparametric regression on the nonparametrically estimated propensity score leads to an efficient estimator. The key reason, which will be detailed later, is the "correction of indexing bias" term in the influence function representation of the estimator, indicating that the bias due to indexing the set of all covariates with a propensity score in the outcome regression model is corrected when one utilizes a nonparametrically-estimated the propensity score rather than the known propensity score.

3.3.3 General Framework of Semiparametric Efficiency Theory – Newey (1994)

For n independent and identically distributed observations $O_i = (Y_i, T_i, X_i)$, suppose the true data generating probability distribution is $P_0 \in \mathcal{P}$, where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a set of possible probability distributions indexed by a parameter θ. That is, $O_i \overset{iid}{\sim} P_0 \in \mathcal{P}$. A parametric model $\mathcal{P}$ is one that is indexed by a finite-dimensional parameter, i.e., $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq R^q\}$. In contrast, in a semiparametric model the parameter θ ranges over an infinite-dimensional parameter space. Unlike the nonparametric model in which the distribution is completely unknown, the semiparametric model has restrictions on a portion of the parameter. The semiparametric model that is of particular interest is one that can be parameterized as $\mathcal{P} = \{P_{(\beta,\eta)} : \beta \in R^q, \eta \in \Theta\}$, where η is the nuisance parameter that ranges over an infinite-dimensional space, and β is the parameter of interest, also called the target parameter.

We now introduce another notation of the target parameter β frequently used in literature: suppose we are interested in estimation and inference for the value $\psi(P)$ of a functional $\psi : \mathcal{P} \to R^q$, which maps the distribution of our observation to the Euclidean parameter space [100]. Consider a parametric submodel indexed by θ: $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta = R^q\} \subset \mathcal{P}$, which goes through P at $\theta = 0$. Estimating $\psi(P)$ is harder than searching within the parametric submode $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta \subseteq R^q\} \subset \mathcal{P}$ and estimating $\psi(P) = \beta$ where $\beta(\theta) = \psi(P_\theta)$. Because we can calculate the Fisher information for estimating $\psi(P_\theta) = \beta(\theta)$,

we can bound the information for estimating $\psi(P)$ using the maximum over the variance bounds of all parametric submodel, as we know that the information would be no bigger than the Fisher information from a parametric model.

According to [72], an influence function is a mean zero function with finite variance that satisfies

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}(O_i) + o_p(1). \tag{3.2}$$

By the central limit theorem, an estimator $\hat{\beta}$ with influence function $\tilde{\psi}$ satisfies that $\sqrt{n}(\hat{\beta}-\beta)$ is asymptotically normal with asymptotic variance equal to $E[\tilde{\psi}\tilde{\psi}^T]$. For an asymptotically linear estimator, the influence function provides information about its asymptotic behavior.

A gradient of a parameter of interest $\beta = \psi(P_0)$ defined at distribution $P_0 \in \mathcal{P}$ is a mean zero function with finite variance that satisfies

$$\frac{\partial \psi(P_\theta)}{\partial \theta}\Big|_{\theta=0} = E[\tilde{\psi}(O)S(O)], \tag{3.3}$$

where $S(O)$ is any score function. A gradient is dependent on both the distribution P_0 and the parameter of interest $\psi(P_0)$, whereas the score function, which sometimes (e.g. when the parameter of interest $\psi(P_{(\beta,\eta)}) = \beta$) turns out to be an influence function up to a normalization, depends only on the distribution $P_{(\beta,\eta)}$. A gradient is closely related to an influence function in the sense that they are the same set of functions for regular and asymptotic linear estimator estimators of $\psi(P)$ [53]. The notation $\frac{\partial \psi(P_\theta)}{\partial \theta}$ is called the pathwise derivative. If Equation (3.3) is satisfied, we call the parameter "pathwise differentiable". To this end, we consider only parametric submodels that are hellinger differentiable.

We note that gradients are not unique. However, the canonical gradient, which has the smallest variance $E[\tilde{\psi}\tilde{\psi}^T]$, is unique. An estimator whose influence function is the canonical gradient achieves the semiparametric efficiency bound, which is the semiparametric analog of the Cramer-Rao lower bound: no regular estimator under the semiparametric model has a smaller asymptotic variance. In addition, its influence function is the efficient influence function.

3.3.4 Semiparametric Efficiency of the ATE via Identification $E\{E[Y \mid T = 1, X]\}$ – Hahn (1998)

In this section, we consider estimation of the ATE. Suppose that the conditional distributions of the potential outcomes are $f_1(y \mid x)$ and $f_0(y \mid x)$, the conditional distribution of the treatment is $p(t \mid x)$, and the distribution of the covariates is $f(X)$, the density function of the observation $o = (y, x, t)$ is

$$[f_1(y \mid x)p(t \mid x)]^t[f_0(y \mid x)p(t \mid x)]^{1-t}f(x). \tag{3.4}$$

A parametric submodel indexed by θ has a density function

$$[f_1(y \mid x;\theta)p(t \mid x;\theta)]^t[f_0(y \mid x;\theta)p(t|x;\theta)]^{1-t}f(x;\theta). \tag{3.5}$$

[39] proved that the semiparametric asymptotic variance bound for estimation of the ATE

$$\beta_* = E[Y(1) - Y(0)] = \int\int yf_1(y \mid x)f(x)dydx - \int\int yf_0(y \mid x)f(x)dydx \tag{3.6}$$

is

$$E\left[\frac{\sigma_1^2(X)}{h(X)} + \frac{\sigma_0^2(X)}{1-h(X)} + (\beta(X) - \beta_*)^2\right] \tag{3.7}$$

where

$$\begin{aligned}
\beta(X) &= E[Y(1) \mid X] - E[Y(0) \mid X] \\
\sigma_1^2(X) &= Var[Y(1) \mid X] \\
\sigma_0^2(X) &= Var[Y(0) \mid X] \\
h(X) &= P[T = 1 \mid X]
\end{aligned}$$

The semiparametric efficiency bound is the variance of the efficient influence function

(EIF)

$$
\begin{aligned}
&\mathrm{EIF}\big(Y=y, T=t, X=x\big) \\
&= E[Y(1) \mid X=x] - E[Y(0) \mid X=x] - \beta_* \\
&+ \frac{t}{h(x)}(y - E[Y(1) \mid X=x]) - \frac{1-x}{1-h(x)}(y - E[Y(0) \mid X=x]) \\
&= E[Y \mid T=1, X=x] - E[Y \mid T=0, X=x] - \beta_* \\
&+ \frac{t}{h(x)}(y - E[Y \mid T=1, X=x]) - \frac{1-x}{1-h(x)}(y - E[Y \mid T=0, X=x])
\end{aligned} \tag{3.8}
$$

As discussed in Section 3.3.1, due to the assumption that $Y(\cdot) \perp T \mid X$, the ATE can be identified by a G-computation parameter

$$\beta_* = E_X\{E[Y \mid T=1, X]\} - E_X\{E[Y \mid T=0, X]\}. \tag{3.9}$$

One could fit two non-parametric regressions

$$E[Y \mid T=1, X=x] = \lambda_1(x) \tag{3.10}$$

$$E[Y \mid T=0, X=x] = \lambda_2(x), \tag{3.11}$$

which are equivalent to modeling

$$E[Y \mid T=t, X=x] = \lambda_2(x) + [\lambda_1(x) - \lambda_2(x)]t. \tag{3.12}$$

Then the ATE can be estimated as

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}_1(X_i) - \hat{\lambda}_2(X_i). \tag{3.13}$$

3.3.5 Semiparametric Efficiency of the ATE via Identification $E\{E[Y \mid X=1, S]\}$ – Hahn (2013)

We consider the causal effect estimator obtained via regression adjustment of the estimated propensity score followed by standardization, which we refer to as three-step estimators. In particular, in the first step a propensity score model is built, in the second step the outcome

is fit against the treatment variable adjusting for the estimated propensity score, and in the third step the predicted pairs of potential outcomes are averaged over the entire population to obtain the mean potential outcomes for marginal comparison that has casual interpretation. We call such an estimator a three-step estimator. Because the estimated propensity score serves as a covariate in the second step for the outcome regression model, such a covariate is referred to as the generated regressor or generated covariate in econometric literature.

A key problem in studying the asymptotic properties of multi-stage semiparametric estimators utilizing generated regressors is how to incorporate the variability of the first stage estimation of the functions that generate the regressors. In particular, the uncertainty of the generated regressor has dual effects on the parameter of interest. The estimated propensity score is an example of generated regressor. On one hand, it influences the predicted value of the potential outcomes, because the outcome model is evaluated at the estimated values of the propensity scores. On the other hand, the outcome model is a function of the propensity score, and the estimation of the function involves the estimated propensity score. In summary, one needs to correctly characterize the effect of the uncertainty of the propensity score on both the estimation and the evaluation of the outcome regression model.

We denote the propensity score, which is essentially a conditional mean function $E[T|X]$ evaluated at a set of covariates X, as $S = h(X)$. Due to the balancing property of S, the ATE is identified by

$$\begin{aligned}\beta_* &= E_S E[Y \mid T = 1, S = h(X)] - E_S E[Y \mid T = 0, S = h(X)] \\ &= E_X E[Y \mid T = 1, h(X)] - E_X E[Y \mid T = 0, h(X)]. \end{aligned} \tag{3.14}$$

[41] derived the influence function for the three-step estimator when the first step is parametric or nonparametric propensity score model. Here we provide a brief summary below.

The first step attempts to fit a regression model to estimate $h(X)$. There are two classes discussed in [41]: (1) a non-parametric model that describe the relationship between t and x as an purely unknown function

$$h_*(x) = E[T \mid X = x] \tag{3.15}$$

with the predicted propensity score denoted as $\hat{h}(X)$; (2) a parametric model

$$h(x;\alpha_*) = E[T \mid X = x] \tag{3.16}$$

where h is known by assumption, e.g. $h(x;\alpha) = (1+\exp(\alpha x))^{-1}$, and α_* is the true coefficients in this model. Denote the estimated propensity score as $h(x;\hat{\alpha})$.

The second step is a non-parametric regression on the propensity score. To emphasize the "true" function, we use "$_*$" to denote the *true* propensity score under a given model, i.e., denote $S_* = h_*(x) = E[T \mid X = x]$ under P, which is equal to $h_*(x)$ or $h(x;\alpha_*)$ depending on model assumption; denote $S_\theta = h_*(x,\theta) = E_\theta[T \mid X = x]$ as the *true* propensity score under P_θ, with $S_\theta \mid_{\theta=0} = S_*$.

The second step models are

$$E[Y \mid T = 1, S = h_*(x)] = g_1(s_*) \tag{3.17}$$

$$E[Y \mid T = 0, S = h_*(x)] = g_0(s_*), \tag{3.18}$$

where $s_* = h_*(x)$, which are equivalent to modeling

$$E[Y \mid T = t, s_*] = g_0(s_*) + [g_1(s_*) - g_0(s_*)]t. \tag{3.19}$$

Note that S is a scalar, whereas the dimension of X can be large, which might cause the curse of dimensionality issue in non-parametric regressions. The true ATE is then

$$\beta_* = E[g_1(h_*(X)) - g_0(h_*(X))]. \tag{3.20}$$

Because we only have the estimated propensity score function $\hat{s} = \hat{h}(X)$ (which can be either $\hat{h}(X)$ or $h(x;\hat{\alpha})$), the non-parametric estimates of $g_1(\cdot)$ and $g_0(\cdot)$ is influenced by the estimated variable $\hat{s}$. To emphasize this, we use $\hat{g}_1(\cdot)$ and $\hat{g}_0(\cdot)$ to denote the estimates of the unknown functions $g_1(\cdot)$ and $g_0(\cdot)$, and have in mind that the functional form of $\hat{g}_1(\cdot)$ and $\hat{g}_0(\cdot)$ is influenced by its argument s. Hence, the estimated conditional means are

$$\hat{E}[Y \mid T = 1, \hat{s} = \hat{h}(X)] = \hat{g}_1(\hat{h}(X)) \tag{3.21}$$

$$\hat{E}[Y \mid T = 0, \hat{s} = \hat{h}(X)] = \hat{g}_0(\hat{h}(X)). \tag{3.22}$$

The third step is called marginalization, which is a population average of predicted potential outcomes $\hat{E}[Y(1)] = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_1(\hat{h}(X_i))$ and $\hat{E}[Y(0)] = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_0(\hat{h}(X_i))$. The estimated ATE is then a comparison of the average potential outcomes, which is

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_1(\hat{h}(X_i)) - \hat{g}_0(\hat{h}(X_i)). \tag{3.23}$$

Depending on how we estimate in the first step, the estimator can take the form of either

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_1(\hat{h}(X_i)) - \hat{g}_0(\hat{h}(X_i)) \tag{3.24}$$

if $h_*(x)$ is estimated non-parametrically, or

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_1(h(X_i;\hat{\alpha})) - \hat{g}_0(h(X_i;\hat{\alpha})) \tag{3.25}$$

if $h_*(x)$ is estimated parametrically.

Now we derive the influence function for the class of three-step estimators by looking at the pathways derivative of the parametric submodels. Note that the ATE β_* satisfies the moment function

$$\begin{aligned} 0 &= E[m\big(X, g_1, g_0, h_*, \beta_*\big)] \\ \text{where } m\big(X, g_1, g_0, h_*, \beta_*\big) &= g_1(h_*(X)) - g_0(h_*(X)) - \beta_*. \end{aligned}$$

Under the parametric submodels $P_\theta(y,t,x)$, we have

$$\begin{aligned} 0 &= E_\theta[m\big(X, g_1, g_0, h_*, \beta_*\big)] \\ &= E_\theta[g_1(h_*(X,\theta),\theta) - g_0(h_*(X,\theta),\theta) - \beta(\theta)] \end{aligned}$$

Thus

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}E_\theta[g_1(h_*(X,\theta),\theta) - g_0(h_*(X,\theta),\theta) - \beta(\theta)] \\ &= \frac{\partial}{\partial\theta}E_\theta[g_1(h_*(X)) - g_0(h_*(X)) - \beta_*] \\ &+ E\big[\frac{\partial}{\partial\theta}\big\{g_1(h_*(X,\theta),\theta) - g_0(h_*(X,\theta),\theta)\big\}\big] \\ &+ E\big[\frac{\partial}{\partial\theta}\big\{-\beta(\theta)\big\}\big] \end{aligned}$$

So we can solve for the pathwise derivative

$$\begin{aligned}
\frac{\partial}{\partial\theta}\beta(\theta) &= \frac{\partial}{\partial\theta}E_\theta[g_1(h_*(X)) - g_0(h_*(X)) - \beta_*] \\
&+ E\big[\frac{\partial}{\partial\theta}\big\{g_1(h_*(X,\theta),\theta) - g_0(h_*(X,\theta),\theta)\big\}\big] \\
&= E[\big(g_1(h_*(X)) - g_0(h_*(X)) - \beta_*\big)S(O)] \text{ (A)} \\
&+ E\big[\frac{\partial}{\partial\theta}\big\{g_1(h_*(X),\theta) - g_0(h_*(X),\theta)\big\}\big] \text{ (B)} \\
&+ E\big[\frac{\partial}{\partial\theta}\big\{g_1(h_*(X,\theta)) - g_0(h_*(X,\theta))\big\}\big] \text{ (C)}
\end{aligned}$$

Here (A) gives the main term of the influence function, which is

$$g_1(h_*(X)) - g_0(h_*(X)) - \beta_*; \tag{3.26}$$

(B) is the adjustment term from estimating the unknown functions g_1 and g_0; (C) is the adjustment term from estimating the unknown function p_*.

We now present several Lemmas on the derivation of adjustment terms (B) and (C).

Lemma 3.1 (Adjustment Term (B))**.** *We have*

$$E\big[\frac{\partial}{\partial\theta}\big\{g_1(h_*(X),\theta) - g_0(h_*(X),\theta)\big\}\big] = E[\big(\frac{T}{h_*(X)}(Y - g_1(h_*(X))) - \frac{1-T}{1-h_*(X)}(Y - g_0(h_*(X)))\big)\cdot S(O)]$$

The adjustment term is

$$\frac{T}{h_*(X)}(Y - g_1(h_*(X))) - \frac{1-T}{1-h_*(X)}(Y - g_0(h_*(X))) \tag{3.27}$$

Proof. The proof is detailed in Appendix B.1. □

Lemma 3.2 (Adjustment Term (C))**.** *We have*

$$E\big[\frac{\partial}{\partial\theta}\big\{g_1(h_*(X,\theta)) - g_0(h_*(X,\theta))\big\}\big] = E[\delta(X)\cdot\big(T - h_*(X)\big)\cdot S(O)],$$

where

$$\delta(X) = (-1)\cdot\big(\frac{E[Y \mid T=1, X] - E[Y \mid T=1, h_*(X)]}{h_*(X)} + \frac{E[Y \mid T=0, X] - E[Y \mid T=0, h_*(X)]}{1-h_*(X)}\big).$$

The adjustment term is

$$(-1)\cdot\Big(\frac{E[Y \mid T=1,X]-E[Y \mid T=1,h_*(X)]}{h_*(X)}+\frac{E[Y \mid T=0,X]-E[Y \mid T=0,h_*(X)]}{1-h_*(X)}\Big)\cdot(T-h_*(X)) \tag{3.28}$$

Proof. The proof is detailed in Appendix B.1. □

Taking a closer look at the adjustment term, we can see that this is the difference between a nonparametric regression on X and on $S = P(X)$, scaled by $\frac{-1}{h_*(X)} \cdot (T - h_*(X))$ or $\frac{-1}{1-h_*(X)} \cdot (T - h_*(X))$.

The following theorem presents the influence function of the three-step parameter.

Theorem 3.1. *The influence function for estimation of the parameter β_* is*

$$\begin{aligned}\Psi(O=o) = {} & E[Y \mid T=1,X=x]-E[Y \mid T=0,X=x]-\beta_* \\ & + \frac{t}{h_*(x)}(y-E[Y \mid T=1,X=x]) - \frac{1-t}{1-h_*(x)}(y-E[Y \mid T=0,X=x]),\end{aligned} \tag{3.29}$$

which is equal to the EIF Equation (3.8).

Proof. The proof is detailed in Appendix B.1. □

The above result implies that using nonparametric regression on a nonparametrically estimated propensity score followed by standardization provides an estimator of the ATE that is semiparametric efficient.

3.3.6 Regularity Conditions for the Asymptotic Normality of $E\{E[Y \mid X=1,S]\}$ via Kernel Smoothing – Mammen (2016) and Regression Splines – Extension of Hahn (2016)

Although [41] (reviewed in Section 3.3.5) has shown the semiparametric efficiency of the general class of three-step estimators that utilizes the generated regressor, they did so by assuming that any required regularity conditions holds for $\sqrt{n}$ consistency and asymptotic

normality but did not specify the conditions explicitly. In this section, we first introduce two commonly used nonparametric smoothing techniques: kernel and spline regression. Then we summarize the regularity conditions detailed in [60] when kernel smoothing is used. We finally extend recent theoretical advances detailed in [40] to provide regularity conditions when one uses regression splines in the three-step estimator.

3.3.6.1 Nonparametric Smoothing Methods

Recall that

$$E[Y \mid T = 1, S = h_*(x)] = g_1(s_*) \tag{3.30}$$

$$E[Y \mid T = 0, S = h_*(x)] = g_0(s_*). \tag{3.31}$$

Equivalently we could model the relationship between the outcome Y, the exposure T, and a single covariate S using a generalized varying coefficient model

$$g\left(E[Y \mid T, S]\right) = \alpha(S) + \beta(S)T \tag{3.32}$$

$$S = E[T \mid X] = h(X) \tag{3.33}$$

where g is the link function in generalized linear models, $\alpha(\cdot)$, $\beta(\cdot)$, and $h(\cdot)$ are three unknown components that are potentially nonlinear functions, which describe the main confounding effect, modification of exposure effect, and exposure proneness respectively. By the definition and property of propensity score, we have $E[Y \mid T = t, S] = E[Y(t) \mid S]$. The conditional mean of potential outcomes in the exposure and control group is thus determined by the two unknown functions, i.e., $E[Y(1) \mid S = s] = g^{-1}\left(\alpha(s) + \beta(s)\right)$ and $E[Y(0) \mid S = s] = g^{-1}\left(\beta(s)\right)$. The ATE from a linear model, i.e., $g(\cdot) = 1$, is $E_S[\beta(S)]$.

Classic parametric models make additional assumptions on $\alpha(\cdot)$ and $\beta(\cdot)$ using a finite number of parameters, whereas nonparametric models relax the form of $E[Y \mid T, S]$ to completely unrestricted functionals $f(S = h(x), T = t)$. There is a tradeoff between reducing bias by flexible modeling to avoid making implausible assumptions, and reducing variation by assuming finite parameters. By flexibly modeling $\alpha(\cdot)$ and $\beta(\cdot)$, while keeping the traditional

parametric modeling of the propensity score γ, we have a semiparametric model which both keeps a certain level of flexibility, and gains efficiency by imposing a scientifically meaningful structure of $\alpha(s) + \beta(s)t$.

There are two major types of nonparametric smoothing technique for estimation of the unknown functions $\alpha(\cdot)$ and $\beta(\cdot)$. One is kernel local-linear or local-polynomial smoothing [28, 49]. The other is regression splines, i.e., piece-wise polynomial functions that are smooth at the joint of each piece, called the knot.

There are two classes of regression splines in general: penalized and unpenalized spline regressions. Smoothing splines and penalized spline regression are two penalized methods that achieve smoothness either by direct penalization on measurement of smoothness, or by penalization on selection of knots.

Polynomial spline regression is a basis-function-based, unpenalized nonparametric regression method, which requires pre-specified placement of knots. After the basis functions are determined, the spline regression reduces to fitting the outcome on the generated basis functions using least square estimation, and thus this method naturally inherits certain asymptotic properties of parametric models. However, under-smoothing, i.e., a potentially larger number of degree of freedom, is often required to obtain a parametric rate of convergence.

One type of spline basis function is the B-spline, which is a nonsingular linear transformation of the polynomial splines to avoid numerical problems associated with multicolinearity. Pre-specifications of the number and placement of knots, as well as the degree of the B-spline are also needed. Quadratic and cubic spline are commonly used and recommended [91, 46, 25, 105].

Compared to polynomial regression, spline regression can introduce more flexibility and produce more stable estimates by increasing the number of knots while keeping the degree fixed. In practice, one can either place more knots at regions where the unknown function varies more rapidly and needs more flexibility, or place knots at the quantiles of the data so that one only need to optimize the number of knots using for example cross validation.

Estimating a conditional mean $E[Y \mid S = s] = f(s)$ has been the focus of studies in

nonparametric regression [73, 104, 20]. However, for causal inference, one is interested in the *marginal* mean $E[Y]$, the estimator of which has different asymptotic properties and is rarely studied to our knowledge. [59] estimated population-level quantities of interest from a generalized partially linear model using kernel estimation, and established theoretical properties of the estimator including normality and semiparametric efficiency. However, to our knowledge, there are few studies on spline regression targeting the marginal mean.

3.3.6.2 Regularity Conditions for Asymptotic Properties of Semiparametric Estimators Using Kernel Regression on Generated Propensity Score – Mammen (2016)

Recently, [60] considered semiparametric estimation with kernel regression on generated regressor, and provided a general approach for asymptotic analysis through stochastic expansions of nonparametric regression estimators. Due to the presence of generated regressor, studying the asymptotic properties of the three-step estimator is no longer standard as in [72, 19]. The work of [60] provided additional required regularity conditions for $\sqrt{n}$ consistency and asymptotic normality. To our knowledge, this is the first paper to give explicit conditions.

Specifically, Proposition 1 of [60] states that:

> Suppose that the regularity conditions given in Appendix B.2 hold, then the ATE estimator $\hat{\beta}$ satisfies $\hat{\beta} \to_d \beta_*$ and $\sqrt{n}(\hat{\beta} - \beta_*) \to_p N(0, E[\Psi(O)^2])$, where β_* is the true value and
>
> $$\Psi(O = (y, t, x)) = E[Y \mid T = 1, X = x] - E[Y \mid T = 0, X = x] - \beta_* + \frac{t}{h_*(x)}(y - E[Y \mid T = 1, X = x]) - \frac{1 - t}{1 - h_*(x)}(y - E[Y \mid T = 0, X = x]),$$
>
> which is equal to the EIF Equation (3.8).

3.3.6.3 Regularity Conditions for Asymptotic Properties of Semiparametric Estimators Using Spline Regression on Generated Propensity Score – Hahn (2016)

Similar to [60], [40] studied two-step sieve M estimation of general semi/nonparametric models, where the second step involves sieve estimation of unknown functions that may use the nonparametric estimates from the first step as covariates, i.e., nonparametrically generated regressors. In this section, we will apply and extend the asymptotic analyses in [40] to the semiparemetric estimator of nonparametric regression on nonparametrically generated propensity score using regression splines in both outcome and propensity score models. We formulate the statistical framework as follows.

Estimation

$$\text{Model: } Y_i = g_0(h(X_i)) + u_i$$

$$T_i = h_0(X_i) + \epsilon_i$$

$$\text{Parameter of Interest: } E[Y(1)] = E[E[Y|S=1,X]] = \int_{\mathcal{X}} g_0(h_0) dP(x).$$

Estimation is $E[Y(0)]$ follows the same asymptotic analysis and we will take the estimation of $E[Y(1)]$ as an example hereafter.

Three-step Estimator of $E[Y(1)]$

Step 1: Estimation of $h(\cdot) = E[T|X=\cdot]$

$$\hat{h}_n = \arg\max_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \frac{-1}{2} [T_i - h(X_i)]^2$$

$$\text{where } \mathcal{H}_n = \{h = R(\cdot)^T \gamma : \gamma \in R^L\} \quad \text{with } L \equiv \dim(\mathcal{H}_n) = L(n),$$

that is, the estimator $\hat{h}_n$ lies in some finite-dimensional sieve space $\mathcal{H}_n$ whose dimension grows to infinity as a function of sample size n.

This optimization problem in fact has a closed form solution which is

$$\hat{h}_n(\cdot) = R(\cdot)^T [(R_n R_n^T)^{-1} R_n S_n] = R(\cdot)^T \hat{\gamma}_n,$$

where $R_n = [R(X_1), \ldots, R(X_n)]$ with $\dim(R_n)$=$L \times n$.

Denote $\varphi(O,h) = \frac{-1}{2}[T - h(X)]^2$, then $\hat{h}_n$ is the maximizer of $\frac{1}{n}\sum_{i=1}^{n}\varphi(O_i,h)$. We assume the identification assumption that $h_0 \in \mathcal{H}$ is the unique solution to $\sup_{h\in\mathcal{H}} E[\varphi(O,h)]$, where $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is some infinite-dimensional separable complete metric space with some norm $\|\cdot\|_{\mathcal{H}}$. Define a neighborhood of h_0 in $\mathcal{H}_n$ as

$$\mathcal{N}_{h,n} = \{h = R(\cdot)^T\gamma_L : \|h - h_0\|_2 \leq \delta_{1,n}\}$$

and an element in $\mathcal{N}_{h,n}$ as

$$h_{0,L}(\cdot) = R(\cdot)^T\gamma_{0,L},$$

as defined in Assumption 1.1 (iii).

Step 2: Estimation of $g(\cdot) = E[Y(1)|\hat{h}(X) = \cdot] = E[Y|T = 1, \hat{h}(X) = \cdot]$

$$\hat{g}_n = \arg\max_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{i=1}^{n} \frac{-1}{2}\frac{T_i}{\hat{h}_i}[Y_i - g(\hat{h}_i)]^2$$

$$\text{where } \mathcal{G}_n = \{g = B(\cdot)^T\beta : \beta \in R^K\} \quad \text{with } K \equiv \ \dim(\mathcal{G}_n) = K(n),$$

that is, the estimator $\hat{g}_n$ lies in some finite-dimensional sieve space $\mathcal{G}_n$ whose dimension grows to infinity as a function of sample size n.

This optimization problem has a closed form solution which is

$$\hat{g}_n(\cdot) = B[h(\cdot)]^T[(\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_nS_n] = B[h(\cdot)]^T\hat{\beta}_n,$$

where $\hat{B}_n = [B(\hat{h}_1), \ldots, B(\hat{h}_n)]^T$, $\hat{h}_i = \hat{h}_n(X_i)$ with $\dim(\hat{B}_n)$=$n \times K$.

Denote $\psi(O,g,h) = \frac{-1}{2}\frac{T}{h}[Y - g(h)]^2$, then $\hat{g}_n$ is the maximizer of $\frac{1}{n}\sum_{i=1}^{n}\psi(O_i,g,\hat{h}_n)$. We assume the identification assumption that $g_0 \in \mathcal{G}$ is the unique solution to $\sup_{g\in\mathcal{G}} E[\psi(O,g,h_0)]$, where $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ is some infinite-dimensional separable complete metric space with some norm $\|\cdot\|_{\mathcal{G}}$. Define a neighborhood of g_0 in $\mathcal{G}_n$ as

$$\mathcal{N}_{g,n} = \{g = B(\cdot)^T\beta_K : \|g - g_0\|_2 \leq \delta_{2,n}\}$$

and an element in $\mathcal{N}_{g,n}$ as

$$g_{0,K}(\cdot) = B(\cdot)^T \beta_{0,K},$$

as defined in Assumption 1.2 (iii).

$$\textbf{Step 3: } \hat{E}[Y(1)] = \frac{1}{n}\sum_{i=1}^{n} \hat{g}_n(\hat{h}_n(X_i))$$

Denote $\rho[g(h)] = \int g(h(x))dP(x)$ the integration over the marginal distribution of X, and $\rho_n[g(h)] = \frac{1}{n}\sum_{i=1}^{n}[g(h(X_i))]$ the empirical average. Then $\hat{E}[Y(1)] = \rho_n[\hat{g}_n(\hat{h}_n)]$, which is a empirical analog of $\rho[\hat{g}_n(\hat{h}_n)] = \int \hat{g}_n(\hat{h}_n)dP(x)$, and both are estimators of $E[Y(1)] = \rho(g_0(h_0))$.

Theorem 3.2. *Under Assumptions 1.1-1.4 in Appendix B.3, we have*

$$\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_0(h_0)] = \frac{1}{n}\sum_{i=1}^{n}\Big\{\triangle_\varphi(O, h_0)[d^*_{\Gamma_n}] + \triangle_\psi(O, g_0, h_0)[d^*_{g_n}]\Big\} + \|v^*_n\|_{sd} o(n^{-1/2}),$$

where

$$\triangle_\varphi(O_i, h_0)[d^*_{\Gamma_n}] = (-1)\frac{E[Y|T=1, X_i] - E[Y|T=1, h_0(X_i)]}{h_0(X_i)}\{T_i - E[Y|T=1, h_0(X_i)]\} + o(n^{-1/2}),$$

$$\triangle_\psi(O_i, g_0, h_0)[d^*_{g_n}] = \frac{T_i}{h_0(X_i)}\{Y_i - E[Y|T=1, h_0(X_i)]\} + o(n^{-1/2}),$$

and

$$\|v^*_n\|^2_{sd} = Var[\frac{1}{n}\sum_{i=1}^{n}\Big(g_0(h_0(X_i)) - \rho[g_0(h_0)] + \triangle_\varphi(O_i, h_0)[d^*_{\Gamma_n}] + \triangle_\psi(O_i, g_0, h_0)[d^*_{g_n}]\Big)].$$

Proof. The proof is detailed in Appendix B.3. □

Corollary 3.1. *Under Assumptions 1.1-1.4 in Appendix B.3, we have that $\hat{E}[Y(1)]$ is asymptotically linear with*

$$\hat{E}[Y(1)] - E[Y(1)] = \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{T_i}{h(X_i)}(Y - E[Y|T=1, X_i]) + E[Y|T=1, X_i] - E[Y(1)]\Big\} + \|v^*_n\|_{sd} o(n^{-1/2}).$$

Consequently, we have

$$\sqrt{n}\Big(\hat{E}[Y(1)] - E[Y(1)]\Big) \to_d N(0, \sigma^2),$$

where $\sigma^2 = Var\{\frac{T}{h(X)}(Y - E[Y|T=1,X]) + E[Y|T=1,X] - E[Y(1)]\}$.

Proof. The proof is detailed in Appendix B.3. □

3.4 Discussion

In this chapter, we first reviewed previous literature on causal inference methods that utilize regression adjustment on the propensity score as a covariate. We considered both linear and nonlinear adjustment and discussed that linear adjustment may not sufficiently control for confounding whereas outcome regression using nonlinear adjustment on the propensity score can serve as an intermediate summary to estimate the mean of potential outcomes over the target population. We then reviewed literature on the theoretical properties of propensity score regression adjustment methods, which can be summarized in the following table (Table 3.2):

Table 3.2: Literature related to propensity score regression adjustment methods

Method	**Observed** S	**Generated Covariate** $\hat{S}$	
	$\frac{1}{n}\sum \hat{\mu}(S)$	$\int \hat{\mu}(\hat{S})dP_X$	$\frac{1}{n}\sum \hat{\mu}(\hat{S})$
General	Newey (1994) [72]	Hahn (2013) [41]	
Kernel $\hat{\mu}(\cdot)$		Mammen (2012) [61]	Mammen (2016) [60]
Spline $\hat{\mu}(\cdot)$		Hahn (2016) [40]	

In this table, a pioneering paper is [41], which introduced a general framework for the study of semiparametric estimators using generated covariate. The authors derived the influence function of the three-step estimator using nonparametric regression on nonparametrically generated propensity score followed by an empirical average to estimate the population-level mean potential outcomes. Their key contribution is in Lemma 3.2, which derived an

additional adjustment term that considers the contribution of the uncertainty from estimation of the propensity score to the influence function. From this adjustment term we can see that one can gain efficiency by regressing on an estimated propensity score rather than a known propensity score. In addition, the estimator using regression on a know propensity score does not achieve the semiparametric efficiency bound. This seemingly counterintuitive phenomenon can be seen via a simple simulation example shown below in Figure 3.1. In each iteration we simulate 1000 observations. In each simulation, we first generate one single covariate $X \sim N(0,1)$, and then a binary exposure Z that follows a logistic regression model $P[Z = 1 \mid X] = \text{expit}\{\sin(\exp(X)) - \cos(X^3)\}$. Lastly we generated a binary outcome Y that follows a logistic regression model $P[Z = 1 \mid X] = \text{expit}\{X + Z + XZ\}$. In the three-step estimation procedure, we first use B-spline regression to estimate a propensity score S, then use B-spline regression with cubic spline to estimate the exposure specific outcome curves $E[Y(1) \mid S]$ and $E[Y(0) \mid S]$. We finally estimate a marginal, population-level odds ratio on the log scale and compare it to the true ATE. In Figure 3.1, when the degree of freedom of B-spline basis functions in the propensity score model is equal to 1, the point correspond to a parametric model regressing Z on X. The grey line corresponds to a know propensity score obtained via plugging X in the model-generating function $\text{expit}[\sin(\exp(X)) - \cos(X^3)]$. We can see that, indeed, regression on the true propensity score may lead to an estimator that has larger bias and variance than the estimators using estimated propensity score. In addition, as the degree of freedom grows larger, the propensity score model gets more nonparametric, and the bias gets smaller until the variability due to too large dimensions starts to dominate. Moreover, increase in degree of freedom also leads to an increase in the variance of the estimates.

The last cell in Table 3.2 that is left in blank is a key contribution of this dissertation, which extended the asymptotic analysis of [40] to study the asymptotic properties of the three-step estimator using spline regression on nonparametrically generated propensity score via spline regression followed by an empirical average. We learned that the three-step estimator is $\sqrt{n}-$consistent and asymptotically linear with influence function that is equal to

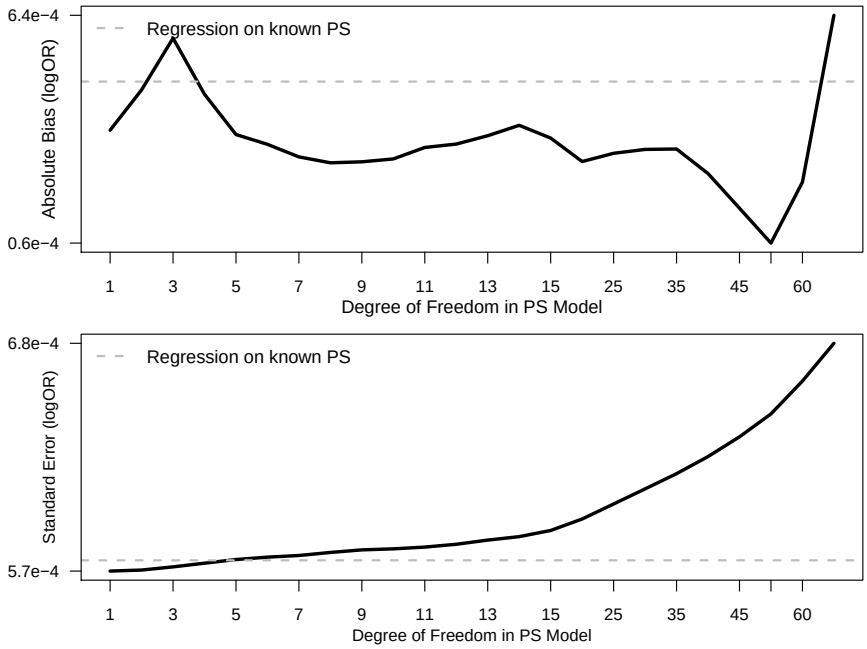

Figure 3.1: Absolute bias and variance of three-step estimators when the propensity score model gets more and more nonparametric.

the efficient influence function, under certain regularity conditions listed in Assumptions 1.1 - 1.4. Particularly, the number of B-spline basis functions L in the propensity score model needs to satisfy that $L \propto n^{\frac{1}{2\rho-1}}$, where $\rho = \frac{\text{number of continuous derivatives of } e(x)=E[Z|X=x]}{\text{dimension of } X}$ measures the smoothness of the function $e(x)$. This indicates that the regularity conditions require an undersmoothed propensity score model, with degree of freedom being slightly larger than the traditional nonparametric rate $n^{\frac{1}{2\rho+1}}$. In contrast, the number of B-spline basis functions K in the outcome regression model needs to satisfy that $\frac{(\log n)^2}{n} L^3 K^5 = o(1)$. For a sufficiently large ρ and sample size n, this indicates that K can be a small value that is smaller than the traditional nonparametric rate. Therefore, for the outcome regression model, the regularity conditions may require an oversmoothed estimate. In practice, a cubic spline with $K = 3$ is good enough. The validity of the list of regularity conditions will be further studied and confirmed in future studies.

Chapter 4

SAFETY SURVEILLANCE AND THE ESTIMATION OF RISK IN SELECT POPULATIONS: FLEXIBLE METHODS TO CONTROL FOR CONFOUNDING WHILE TARGETING MARGINAL COMPARISONS

4.1 Introduction

The increasing availability of electronic health record (EHR) and claims data has created the potential for population scale observational biomedical research. One transformative national effort is the Food and Drug Administration's (FDA) Sentinel Initiative that aims to monitor and evaluate the safety of all regulated medical products [10]. For example, a recent observational cohort study from the FDA Sentinel Initiative compares the effect of Angiotensin-Converting Enzyme Inhibitors (ACEI) and Beta Blockers (BB), two blood pressure control medications, on incidence of angioedema in the first 30 days after starting treatment [69]. The Sentinel system utilizes a distributed data network that provides access to electronic healthcare data for approximately 193 million patients. Data of this quantity is powerful for safety studies because it represents a broader population than typically enrolled in clinical trials, and it enables detection of potential safety signals for less common adverse outcomes.

While the use of such large-scale healthcare data presents numerous opportunities for postmarketing safety research, there are also many inferential challenges. One key challenge is the need to control for a large number of potential confounders in EHR data, which is further complicated by the fact that an adverse event is often extremely rare. When the outcome is rare, using regression adjustment with a large number of covariates can result in model fitting issues such as non-convergence or substantial instability. Flexible nonparametric regression

is even more challenging due to the well-known curse of dimensionality [43]. In contrast, if there is sufficient uptake of both the new medical product of interest and the control, a comparator medical product, fitting a propensity score (PS) model with a large number of covariates may be more feasible.

Because use of propensity scores can provide both control of confounding and dimension reduction, it is attractive to consider propensity score methods in the postmarket surveillance setting [87]. One such propensity score method is the direct regression adjustment of the propensity score as a covariate in an outcome regression model. To estimate marginal, population-level contrasts that are the central focus of causal inference, the use of a regression model is often considered as the intermediate summary that is then used in a final standardization step[1], which takes the empirical average of the pair of predicted risks over the entire population under hypothetical exposure and control conditions.

Propensity score regression adjustment coupled with standardization has not been well-studied and potentially underused. First, [4] and [3] have shown in simulation studies that regression adjustment on the propensity score can result in a biased effect estimate. However, in these studies, the propensity score was adjusted as a linear term, which may not fully capture the relationship between the outcome and the propensity score. Therefore, previously observed bias could be due to model misspecification or residual confounding rather than the validity of the propensity score regression adjustment method [37, 102]. Second, statistical inference can be challenging because using the estimated propensity score as a covariate introduces additional uncertainty from the preliminary propensity score model, and this variability needs to be accounted for in the ultimate estimation of standard errors [41].

Despite the benefits and popularity of propensity score methods in postmarket surveillance, there have been few studies comparing the performance of propensity score methods in the setting of rare outcomes with many confounders. [29] compared propensity-based es-

[1]In causal inference and missing data literature, the standardization procedure has also been named G-computation [80, 81, 96], marginalization [30, 34], partial or full means [71], marginal integration [98, 55], full imputation [21, 67], or substitution estimator [99].

timators of the marginal relative risk mimicking confounders from EHR studies and assumed a relatively high event rate of 5% yielding 250 events among $5,000$ patients. However, in postmarket surveillance, we often encounter quite rare events with a much smaller event rate and a larger population, such as 50 events among $10,000$ patients at an event rate of 0.5%. In addition, there was no statistical inference or variance estimation considered in their study.

In this paper, we focus on developing a propensity-based estimator for analyzing data with rare binary adverse events. In particular, we characterize a simple three-step method that utilizes the generalized varying coefficient model to provide flexible nonlinear adjustment of the estimated propensity score in an outcome regression model, followed by standardization to estimate a marginal causal effect in a select population. In particular, using this methodology, our inferential procedure targets common population-level contrasts that have a simple and direct causal interpretation, such as the risk difference, the risk ratio, or the odds ratio. In addition, we provide a direct and simple variance estimator for the population-level drug effect through adoption of a unified influence function representation that fully accounts for the uncertainty from propensity score estimation, outcome modeling, and standardization. We conduct a realistic simulation study by mimicking real data from the FDA Sentinel Initiative study comparing the effect of ACEIs and BBs on incidence of angioedema in the first 30 days. We look at both the performance of different propensity score methods and the validity of our variance estimate.

Our paper is organized as follows. In Section 2 we detail flexible propensity score adjustment and provide an empirical estimator of the variance. Section 3 presents an overview of existing methods for causal inference using propensity scores which estimate the exposure effect targeting a specific population. In Section 4, we conduct simulation study to compare flexible regression adjustment of the propensity score with existing propensity score methods in the context of causal inference. We close with a discussion in Section 5.

4.2 Flexible Regression on the Propensity Score

4.2.1 Notation in the Formal Language of Causation

Causation is traditionally inferred by the difference in outcomes when all circumstances are the same except for one factor which condition was changed. For each subject in a target population, there is a pair of variables $\big(Y(1),\ Y(0)\big)$ that characterizes the potential outcomes that would have been observed under exposure and control, respectively. The main goal of causal inference is to provide a comparison of the population-level averages between the two potential outcomes. For example, a common causal effect is the average treatment effect, defined as $\text{ATE} = E[Y(1)] - E[Y(0)]$.

In this paper, we focus on statistical inference on the average treatment effect. We denote Z as the binary exposure, taking on value 1 (exposed) or 0 (unexposed). We define the observed outcome as $Y = Y(1)$ if $Z = 1$, and $Y = Y(0)$ if $Z = 0$. Since only one outcome per subject can be observed at a time according to the subject's exposure group, thus, $Y(1)$ is unobserved in the control group and $Y(0)$ is unobserved in the exposure group. In observational studies, the exposed and unexposed have systematic differences in their characteristics $\mathbf{X}$, which are potentially associated with the outcome. These patient characteristics are referred to as confounding variables. Consequently, the distribution of the observed $Y(1)$'s in the exposure group can not represent the distribution of $Y(1)$'s in the entire population.

To mitigate this issue, [87] proposed the strongly ignorable exposure assignment assumption $\big(Y(1),\ Y(0)\big) \perp Z \mid \mathbf{X}$, which implies conditional independency of the potential outcomes and exposure given measured confounders. This assumption allows one to estimate the mean of potential outcomes $Y(1)$ using only the observed portion of $Y \mid Z = 1$ by restricting to a stratum of $\mathbf{X}$, i.e., $E[Y(1) \mid \mathbf{X}] = E[Y \mid Z = 1, \mathbf{X}]$.

In addition, [87] defined the propensity score as the probability of being exposed given the subject's characteristics, i.e., $S = P(Z = 1 \mid \mathbf{X})$. The propensity score is a one-dimensional balancing score: within a stratum of S, the covariates are similar between both exposure and control groups. Thus, the observed portion can represent the entire population in terms

of the mean outcomes, i.e.,

$$E[Y(1) \mid S] = E[Y \mid Z = 1, S]. \tag{4.1}$$

[87] proposed to substitute the set of covariates $\mathbf{X}$ with the propensity score S in an outcome regression model, which allows for both control of confounding and dimension reduction. This approach could potentially gain efficiency in finite sample and is particularly useful when the outcome is rare.

4.2.2 Three-Step Estimation Procedure

In this section, we detail a three-step method that flexibly adjusts for confounding using the propensity score and then standardizes to a target population for causal comparison. Compared to direct covariate adjustment, which will be discussed in Section 4.3, we have replaced the set of covariates with a function of the propensity score in the model to reduce the dimensionality of the covariates while attempting to minimize model assumptions. We further derive variance estimates that incorporate the uncertainty due to estimation of the propensity score in our model.

At the first step, we estimate the propensity score based on a parametric model such as the logistic regression. Recall that we take Z as the binary exposure, such as being exposed to a drug or product. Thus, the propensity score, i.e., the probability of being exposed can be predicted as

$$\hat{S} = \hat{P}[Z = 1 \mid \mathbf{X}] = \left[1 + \exp(-\mathbf{X}\hat{\gamma})\right]^{-1},$$

where $\hat{\gamma}$ a set of estimated coefficients obtained from fitting a logistic regression.

At the second step, we fit a flexible model of the outcome curves using the estimated propensity score as the covariate. We consider the varying coefficient model proposed by [45]:

$$g\left(E[Y \mid Z, \hat{S}]\right) \;=\; \beta(\hat{S})Z + \alpha(\hat{S}), \tag{4.2}$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are unknown and potentially nonlinear functions that balances the exposure arms through the estimated propensity score in order to control for confounding, and $g(\cdot)$ is a known link function. For binary outcomes, $g(\cdot)$ is often the logit link function. When $\beta(\cdot)$ is equal to a constant β, which can be interpreted as the conditional exposure effect, the varying coefficient model reduces to a partially linear model [44]. Combining (4.1) and (4.2), we can see that the varying coefficient model allows us to estimate two curves that predict the means of potential outcomes given patient's treatment proneness:

$$\begin{aligned} E[Y(1) \mid \hat{S}] = E[Y \mid Z = 1, \hat{S}] &= g^{-1}\big[\beta(\hat{S}) + \alpha(\hat{S})\big], \\ E[Y(0) \mid \hat{S}] = E[Y \mid Z = 0, \hat{S}] &= g^{-1}\big[\alpha(\hat{S})\big]. \end{aligned}$$

To estimate the nonlinear functions $\alpha(\cdot)$ and $\beta(\cdot)$, we apply the polynomial spline regression [24, 23]. A spline is a piece-wise polynomial function that is smooth at the joint of each piece, referred to as the knots. Any spline function on a given set of knots can be expressed as a linear combination of B-splines. We generate a set of B-spline basis functions with K knots, $\mathbf{B}(S) = [B_1(S), \ldots, B_K(S)]$, and then fit the outcome on the basis functions and the exposure indicator. The pair of mean potential outcomes under being exposed and unexposed, is predicted as

$$\begin{aligned} \hat{E}[Y(1) \mid \hat{S}] &= g^{-1}\big[\mathbf{B}(\hat{S})\hat{\boldsymbol{\beta}} + \mathbf{B}(\hat{S})\hat{\boldsymbol{\alpha}}\big], \\ \hat{E}[Y(0) \mid \hat{S}] &= g^{-1}\big[\mathbf{B}(\hat{S})\hat{\boldsymbol{\alpha}}\big], \end{aligned}$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the coefficients of the B-spline basis functions.

As discussed in Section 4.2.1, causal inference is a comparison of the population-level averages. Thus, the outcome regression model should often be considered as the intermediate summary that is then used in a final standardization step to yield marginal, population-level contrasts. One popular approach is to take the empirical averages of the predicted risks resulting from creating a pair of predictions for each patient as if they were exposed to each of the two different drugs regardless of their actual exposure condition. Such a procedure is called model-based standardization, which we will refer to generally as standardization

[32, 57]. Therefore, at the third step, we take the empirical average of the estimated potential outcomes over the entire target population:

$$\begin{aligned}\hat{E}[Y(1)] &= \frac{1}{n}\sum_{i=1}^{n}\hat{E}[Y_i(1) \mid \hat{S}_i] = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\big[\mathbf{B}(\hat{S}_i)\hat{\boldsymbol{\beta}} + \mathbf{B}(\hat{S}_i)\hat{\boldsymbol{\alpha}}\big],\\ \hat{E}[Y(0)] &= \frac{1}{n}\sum_{i=1}^{n}\hat{E}[Y_i(0) \mid \hat{S}_i] = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\big[\mathbf{B}(\hat{S}_i)\hat{\boldsymbol{\alpha}}\big].\end{aligned}$$

With the pair of population-level averages of potential outcomes, we can now make simple comparisons that have explicit causal interpretations. For example, for a binary angioedema adverse event outcome, we have a pair of mean risks denoted by $\hat{p}_1 = \hat{E}[Y(1)]$ and $\hat{p}_0 = \hat{E}[Y(0)]$, which are the risks of angioedema among the full population in need of high blood pressure control medications (a combination of the ACEI and BB groups), had they taken ACEI ($\hat{p}_1$) or BB ($\hat{p}_0$). We plug in the estimated mean risks to estimate the parameter of interest such as the risk difference $\hat{\mathrm{RD}} = \hat{p}_1 - \hat{p}_0$, the relative risk $\hat{\mathrm{RR}} = \hat{p}_1/\hat{p}_0$, or the odds ratio $\hat{\mathrm{OR}} = \big[\hat{p}_1/(1-\hat{p}_1)\big]\big/\big[\hat{p}_0/(1-\hat{p}_0)\big]$.

4.2.3 Variance Estimation

To derive the variance of the proposed estimator, we need to incorporate the variability due to estimation of the propensity score, which impacts both the estimation of the nonlinear function $\hat{\alpha}(\cdot)$, and the evaluation of the function $\hat{\alpha}(\hat{S})$ when plugging in $\hat{S}$. [41] studied the asymptotic variance of a class of estimators that employ covariates estimated from a preliminary regression, such as the use of the estimated propensity score. Motivated by [41], we now derive an estimate of the variance for the proposed estimator.

To this end, we introduce the notion of influence function (IF), which measures the influence of changing one data point from the n samples on an estimator. The variance of the estimated influence function for a particular parameter has been used to create an asymptotic variance estimator for the parameter [35]. Under some regularity conditions, it turns out that nonparametric regression on the estimated propensity score $\hat{S}$ with standardization has the same influence function as nonparametric regression on all covariates $\mathbf{X}$ with standardization

[41].

More specifically, let

$$\hat{\text{IF}}_1 = \hat{E}[Y(1) \mid \mathbf{X}] - \hat{p}_1 + \frac{Z}{S}(Y - \hat{E}[Y(1) \mid \mathbf{X}])$$

and

$$\hat{\text{IF}}_0 = \hat{E}[Y(0) \mid \mathbf{X}] - \hat{p}_0 + \frac{1-Z}{1-S}(Y - \hat{E}[Y(0) \mid \mathbf{X}]),$$

where $\hat{p}_1 = \hat{E}[Y(1)]$ and $\hat{p}_0 = \hat{E}[Y(0)]$. By an application of the delta method, one can show that the IF for risk difference (RD) is $\text{IF}_{\text{RD}} = \text{IF}_1 - \text{IF}_0$; the IF for log of the risk ratio (RR) is $\text{IF}_{\text{logRR}} = \frac{\text{IF}_1}{\hat{p}_1} - \frac{\text{IF}_0}{\hat{p}_0}$; and the IF for log of the odds ratio (OR) is $\text{IF}_{\text{logOR}} = \frac{\text{IF}_1}{\hat{p}_1(1-\hat{p}_1)} - \frac{\text{IF}_0}{\hat{p}_0(1-\hat{p}_0)}$. The variance estimator of a parameter of interest, e.g. the risk difference, is then estimated as

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \hat{\text{IF}}^2_{\text{RD},\, i} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\text{IF}}_{1i} - \hat{\text{IF}}_{0i})^2.$$

For construction of confidence intervals and testing, we use the standard error estimated as $\hat{\sigma}/\sqrt{n}$. See the supplementary material for details on derivation of the influence functions.

The IFs involve estimates of the conditional means of the outcomes $\hat{E}[Y_i(1) \mid \mathbf{X}_i]$ and $\hat{E}[Y_i(0) \mid \mathbf{X}_i]$. For our proposed method of regression on the propensity score, estimating the IFs introduces an extra step of estimating the mean outcome conditional on all covariates. In practice, we propose to use the mean outcome conditional on the propensity score instead, which is computed in our estimation procedure. That is, we estimate the influence functions as

$$\widetilde{\text{IF}}_{1i} = \hat{E}[Y_i(1) \mid \hat{S}_i] - \hat{p}_1 + \frac{Z_i}{S_i}(Y_i - \hat{E}[Y_i(1) \mid \hat{S}_i])$$

and

$$\widetilde{\text{IF}}_{0i} = \hat{E}[Y_i(0) \mid \hat{S}_i] - \hat{p}_0 + \frac{1-Z_i}{1-S_i}(Y_i - \hat{E}[Y_i(0) \mid \hat{S}_i]).$$

This would generally yield a conservative variance estimate [41].

We applied this variance estimator to all approaches that utilize standardization following regression adjustment in simulation and application studies, such as regression on covariates (detailed in Section 4.3.3) and linear adjustment of the propensity score. When the outcome

regression on all covariates is used, we plug in IF to estimate the variance; when the outcome is fit on the propensity score, we use $\widetilde{\text{IF}}$ instead.

4.3 A Review of Existing Propensity Score Methods for Binary Outcomes

In this section, we briefly review existing methods that can estimate a pair of population-level mean risks for binary outcomes, which will be plugged in to estimate a population-level risk difference, risk ratio, or odds ratio. Note that all such population-level comparisons have causal interpretation. We focus on methods that utilize the propensity score. In Section 4.4, we will compare these proposals to our proposed flexible regression on the propensity score with standardization.

4.3.1 Propensity Score Matching

The propensity score matching method selects a subpopulation that includes matched sets of exposed and unexposed subjects sharing similar propensity scores, with the goal of mimicking a population from a randomized study. In general, we match each exposed subject to M unexposed subjects and then estimate the exposure effect as if the matched population was observed from a randomized study. Note that the matched subpopulation contains subjects with characteristics similar to the exposed arm. Thus the estimated causal effect is in fact the average treatment effect on the treated population (ATT).

In practice, applying propensity score matching involves decisions on the value of M and a caliper that defines the tolerance of the difference in propensity scores for a matched pair, as well as the sampling method, i.e., with or without replacement. Simulation studies showed that increasing M tended to increase the bias but decrease the variability [6]. [22] provided suggestion on specifying the caliper. For matching with replacement, a pseudo-population that is closest to the exposed population is generated. However, it is hard to interpret the result, and requires accounting for duplicated observations. In general, because the matching procedure does not involve the outcome, one could try multiple values of M and caliper and select the best matched dataset according to covariate balance.

For simplicity in the simulation study and application we did not take into account the variability of the propensity score when estimating the variance, because matching methods are computationally intensive particularly in the postmarket surveillance setting and it was not very feasible to use common approaches like the bootstrap. Omitting the uncertainty in the propensity score would likely increase the power for the matching methods artificially.

4.3.2 Augmented Inverse Probability of Treatment Weighting

Another way to achieve balance in the population is to reweight every subject to create a pseudo-population in which every exposed/unexposed pseudo-subject has equal possibility of being exposed/unexposed. Such a pseudo-population is representative of one from a randomized study. This is called the inverse probability of treatment weighting (IPTW) [82]. A commonly used weight is the inverse of the propensity score, that is, to use $\frac{1}{S}$ if subject is exposed and $\frac{1}{1-S}$ if subject is unexposed. A well-known problem with IPTW is the instability from inverting the estimated propensity score. Stabilized weights have been proposed [82]. Truncation using either a pre-specified threshold or a quantile is also widely used in practice [76, 77], which will be implemented in our simulation study in Section 4.4.

Simple IPTW requires that the propensity score must be correctly specified. To relax this assumption, the Augmented IPTW (AIPTW) approach was proposed, which uses a combination of the propensity score model and the outcome regression model [83, 8]. It has also been referred to as the doubly robust estimator because it consistently estimates the truth when either the propensity score model or the outcome regression model is correctly specified. However, since the estimation of the outcome model is required, there may be convergence issues in the rare event setting. In Section 4.4 we only show the AIPTW method, but not the IPTW, since actual performance was very comparable in our simulation setting. This is likely due to the extremely rare event setting where a bias correction from the outcome regression model using predicted risks is small.

4.3.3 Direct Covariate Regression Followed by Standardization

As outlined in Section 4.2, standardization following flexible propensity score adjustment is a viable approach to estimate causal effects. Another common alternative is to use standardization following direct adjustment for confounders in the outcome regression model. Specifically, first we build an outcome regression model with both exposure and all confounders in the model $g\left(E[Y \mid Z, \mathbf{X}]\right) = \beta Z + \mathbf{X}\boldsymbol{\alpha}$. Then we use standardization, which is done by simply taking the average of the predicted potential outcomes for all individuals in the population as outlined in Section 4.2.2. The estimated causal effect is a comparison of the marginal, population-level mean risks, obtained by plugging in the marginalized means into risk difference, risk ratio, or odds ratio. Compared to Section 4.2.2, there is no estimation of the propensity score, and all confounders are directly adjusted for in the outcome regression model. Therefore, application of this method in the rare event setting may have model fitting issues and instability.

4.4 Simulation

In this section, we perform extensive simulation studies to investigate the performance of our proposed method and the existing methods outlined in Sections 4.2 and 4.3. We consider estimating a marginal OR in the observational surveillance with rare outcomes setting, since it is the most common estimand of interest in observational cohort studies for binary outcomes. Moreover, in the rare event setting, the marginal OR is approximately the relative risk. Our simulation study will mimic real data from the FDA Sentinel Initiative study comparing the effect of angiotensin-converting enzyme Inhibitors (ACEI) and beta blockers (BB) on incidence of angioedema in the first 30 days [69]. Further details of the study are outlined in the simulation setting in Section 4.4.1 and the real data application in Section 4.5.

We consider the following marginal OR estimators: (1) 1-1 matching on the propensity score without replacement (using 0.2 times the standard deviation of the observed outcomes

as the caliper); (2) Augmented IPTW (also referred to as the doubly robust estimator), with parametric models for exposure and outcome both adjusting for main terms of all covariates, with stabilized propensity score by trimming at 5% tail; (3) Regression on covariates with standardization; (4) Regression on main term of the propensity score (linear adjustment) with standardization; (5) Regression on propensity score deciles with standardization; and (6) Flexible regression of the propensity score using B-spline basis functions (here we used cubic spline with one inner knot) with standardization. For simplicity, we use the partially linear model introduced in Section 4.2.2 in the estimation of (5) and (6). That is, we estimate a nonlinear function $\alpha(S)$ using either step functions in (5) or spline regression in (6) to flexibly control for confounding, while imposing a marginal structure that $\beta(S)$ is equal to a constant β.

The first three methods are standard approaches used for estimating a marginal OR, although one-to-one matching without replacement may target at the ATT as discussed in Section 4.3.1. The simulation study of [97] found that in the observational surveillance with rare outcomes setting, regression adjustment on potential confounders performed better than matching in terms of power. Our study will further evaluate the performance of regression adjustment on the propensity score targeting marginal comparisons using methods (4) - (6) to provide guidance on method selection for control of confounding. Method (4) has been shown to be biased in other simulation studies [4, 3] and therefore we were interested to assess for our simulation scenario if these findings still hold, as well as to investigate the potential for bias correction via flexible modeling. Method (5) essentially fits a piecewise constant function or step function, whereas method (6) fits a curve. Both method (5) and (6) provides a more flexible nonlinear function of the propensity score, and may reduce the residual confounding from regression on a simple linear term in method (4). Method performance was assessed in terms of bias and standard error on the log OR scale, as well as type I error and power. For methods (2)-(6), we estimated the variance of the marginal OR following the proposal in Section 4.2.3. For matching methods we calculated the bias by comparing the estimate to both the true ATT and the true ATE.

4.4.1 Simulation Setting

We generate a realistic population of 100,000 subjects mimicking data from the ACEI and BB example. Specifically, there are nine binary clinically relevant covariates (NSAIDs (Nonsteroidal anti-inflammatory drugs), aspirin, ORAL-CS (optimizing recovery after laparoscopic colon surgery), allergic reaction, diabetes, heart disease, Ischemic HD (heart disease), inpatient hospitalization, and gender) and one categorical variable which is age category with four levels, corresponding to three dummy variables (binary indicators). See Table 4.1 for the prevalence (prev) of each confounder.

To simulate the ACEI and BB dataset, we used the following procedure and generated:

(1) Binary and categorical covariates $\mathbf{X}$ that have the same mean and pairwise covariance as the real data, yielding correlated confounders;

(2) A binary exposure Z (ACEI = 1 and BB = 0) generated based on a logistic regression on the covariates (the propensity score model, see Table 4.1), using coefficients observed from fitting the real data. For all cases, we hold the exposure prevalence the same as the real data at 69% for ACEI.

(3) A pair of binary potential outcomes $(Y(1), Y(0))$ (angioedema within 30 days under exposure and control for the same subject) based on a logistic regression on the exposure and covariates (the outcome regression model, see Table 4.1), using the coefficients observed from fitting the real data. We used the pair of potential outcomes to calculate the true ATE and ATT. For methods comparison we used the observed outcome only, which is $Y = ZY(1) + (1 - Z)Y(0)$. For all cases, we hold the event rate in the control group (BB group) the same as the real data, which is equal to 0.03%.

In addition, we increased the strength of confounding by scaling up the coefficient in the propensity score model (multiply coefficients on the logOR scale by 1.5), while still holding the exposure prevalence and the baseline event rate the same (see Table 4.1 stronger propensity). We also allowed the propensity score model or the outcome regression model to include interaction terms between age and diabetes, to look at cases when the methods

misspecify one of the models by missing the interactions in the model (see Table 4.1 adding Interactions). We generated the potential outcomes under both the null model ($\log(\text{OR}) = 0$, i.e., no elevated risk of angioedema due to treatment with ACEI), and the alternative using coefficient $\log(\text{OR}) = 1.1$ ($\text{OR} = 3$) which represents moderate to strong exposure effect. In each scenario assessed we used 4000 simulated datasets.

The outcome event rate in the control group is 0.03%, which yields about 30 angioedema incidences among the $100,000$ subjects. We have 12 confounder indicators in the model including nine binary variables and three indicators for age categories. Note that since the exposure rate for ACEI (69%) is more common than BB, for the matching methods the estimate of interest may be closer to the effect amongst the BB population. We have also simulated the case when exposure rate is 20%, which is more commonly seen for new medications. For simplicity, we use the abbreviation PS to refer to the propensity score hereafter.

4.4.2 Results

Table 4.2 shows the bias and standard error on the log(OR) scale, as well as the type I error and power when both the PS model and outcome regression model are correctly specified. We showed performance under two scenarios: an exposure rate of 69% and 20%. Each scenario was further tabulated by null and alternative drug effect, as well as original and stronger coefficient in the propensity score model which corresponds to strength of confounding under the alternative.

When the exposure rate is 69%, which is the observed value in the real data, simple linear adjustment of the PS had a notable increase in bias under the alternative ($\log(\text{OR}) = 1.1$, $\text{OR} = 3$), which can be as much as 34 times more than its bias under the null. There was no such substantial elevation under the null because the covariates are essentially precision variables, and thus there is no residual confounding from insufficient adjustment. PS B-Splines with standardization was able to correct the bias in linear adjustment of the PS, and had relatively small bias and standard error across all scenarios. PS deciles with

standardization had the smallest standard error, although the performance in terms of bias was not stable. All standardization following regression methods had similar type I error and power. The valid type I error observed in all standardization methods showed that the direct estimation of variance we proposed is a reliable approach for inference which does not require computationally intensive methods such as bootstrap.

When the exposure rate is 20%, which is further from 50% compared to 69%, matching had smaller bias than other methods, but it targets a different parameter (ATT) so comparison with other methods is more difficult. Particularly, comparing matching estimate to the ATE, the bias was in general larger than the standardization methods. Therefore, when ATE is of interest, using one-to-one matching without replacement is expected to be biased unless the true exposure effect is zero. In addition, matching had lower power and higher standard error than other methods. Note that the matched sample has at most 40% of the population which may be the reason that we observed instability in matching. Again, although simple linear adjustment of the PS gained efficiency and had the smallest standard error, it had substantial increase in bias under the alternative due to insufficient adjustment for confounding. PS B-spline with standardization still had small bias without loss of efficiency, and outperformed all other methods with the least bias and most power under the alternative with strong confounding effect. The AIPTW had the largest bias, inflated type I error and lower power, which is an evidence of sensitivity to inverting an estimated PS that may be closer to zero [84]. Moreover, under the null with strong propensity effect, all standardization methods had an increase in bias as well as inflated type I errors due to the highly imbalanced treatment arms.

Table 4.3 shows method performance when the PS model is misspecified by not including the interaction terms. Table 4.4 considers the scenario when the outcome regression model is similarly misspecified by ommission of interaction terms. In both tables, in general we observed slightly increased bias, standard error, and type I error, as well as decreased power compared to Table 4.2. In Table 4.3, when the propensity score model is misspecified, matching had increased bias under the alternative. Regression on the PS (linear PS, PS

B-spline, and PS deciles) with standardization had slightly increased bias and loss in power, particularly under strong confounding effect, but magnitude of the bias and power was similar to most of the other methods evaluated. Regression on PS deciles had unstable performance with potentially decreased bias under the original confounding effect with an exposure rate of 20%. The AIPTW method was more sensitive to misspecification in propensity score model than in outcome regression model, due to inversion of the PS which inflates the perturbation. Regression on covariates with standardization was less sensitive to misspecification in the PS, but had generally larger bias and type I error when the outcome regression model was misspecified as is observed in Table 4.4. In addition, in Table 4.4 when the outcome regression model is misspecified, all methods had a substantial increase in standard error. In both Table 4.3 and Table 4.4, we again see that PS B-spline regression with standardization corrected the substantially elevated bias from adjusting for the PS as a linear term.

Our simulation indicated that the bias of regression on the PS observed in previous simulation studies [4, 3] could be due to residual confounding from simple linear adjustment. Based on our consistent observation that flexible adjustment of PS reduces bias from insufficient linear adjustment and potentially outperforms traditional methods, it is promising in safety surveillance with rare outcomes to estimate a balancing score that sufficiently controls for confounding and reduces dimension to allow for flexible outcome modeling. In addition, the valid type I error and high power showed that our proposed direct estimation of variance is a fast and valid approach for inference.

4.5 Application to the ACEI and Angioedema Study

In this section, we analyze a subset of data obtained from an observational cohort study using EHR data from 2008-2012 at Group Health Cooperative, a managed healthcare system in Washington State that is part of the FDA Sentinel's network [69]. The goal of this evaluation is to compare the effect of angiotensin-converting enzyme Inhibitors (ACEI) and beta blockers (BB), which are two medications used to control high blood pressure, on incidence of angioedema in the first 30 days after starting either medication. There is a known

elevated risk amongst those who take ACEIs relative to BBs for incidence of angioedema especially early after initial drug exposure [89].

Our cohort includes 31,269 prescribed to ACEI and 15,025 to BB. Amongst those prescribed to ACEIs, 49 subjects had an angioedema event (0.157%), and 5 had an angioedema event (0.033%) amongst BB prescribers yielding an unadjusted OR of 4.72. We reanalyze this data set with all of the methods described in the previous sections. For the analysis, we include all of the following potential confounders: NSAIDs (Nonsteroidal anti-inflammatory drugs), aspirin, ORAL-CS (optimizing recovery after laparoscopic colon surgery), allergic reaction, diabetes, heart disease, Ischemic HD (heart disease), inpatient hospitalization, and gender) and one categorical variable which is age (categories: 18-44, 45-54, 55-64, and 65-99). This example was also mimicked for the simulation study presented in the Section 4.4.1. Details of the confounders including the prevalence and relationship to the exposure and the outcome were shown previously in Table 4.1.

We present in Table 4.5 results of applying the different adjusted marginal OR methods to the ACEI and BB cohorts. The methods considered are the same as the ones evaluated in our simulation studies in Section 4.4.1: (1) 1-1 matching on the propensity score without replacement; (2) Augmented IPTW, with parametric models for exposure and outcome, both adjusting for all covariates with trimmed propensity score using 5% tail as the threshold; (3) Regression on covariates with standardization; (4) Regression on linear propensity score adjustment with standardization; (5) Regression on propensity score deciles with standardization; and (6) Flexible regression of the propensity score using B-spline basis functions (here we used cubic spline with one inner knot) with standardization.

All of the methods find a statistically significant association in increased risk for angioedema when the entire population is treated with ACEI, compared to when the entire population is treated with BB. The estimated marginal OR is approximately 5.5 across all methods. Note that propensity score matching estimates an OR in a different population: amongst the group of patients that are actually treated with ACEI and found a viable propensity score match with a BB patient, and thus the value is potentially different. In the

entire population, the average adjusted risk of angioedema in 30 days under ACEI treatment is around 0.17% (17 per 10,000) whereas the average adjusted risk of angioedema in 30 days under BB is around 0.03% (3 per 10,000). Among all confounding control methods, our proposed method of standardization following regression on propensity score B-spline has the smallest standard error and therefore indicating the most efficient method in this example. Also note that propensity score matching has the largest standard error which may be expected since it only uses a subset of the cohort which is not optimal especially in this very rare event setting.

4.6 Discussion

In this paper, we have shown that there is a great potential in using standardization following regression adjustment of the propensity score to estimate a marginal, causal effect in a select population for rare binary outcomes. While the propensity score is sufficient in balancing the confounders between exposure groups, adjusting the propensity score as a linear covariate in the outcome regression model can result in bias [4, 3]. By fitting flexible spline function of the propensity score in the outcome regression model, our proposed method is a fast and simple correction of this bias. In addition, our proposed direct estimation of the variance is a fast and reliable approach for performing inference compared to computationally more extensive approaches such as the bootstrap procedure.

Our simulation studies have shown that flexible adjustment of the propensity score in an outcome regression model with standardization resulted in less bias without loss of efficiency. Moreover, it performs equivalently and often better than the existing methods when the propensity score model is correctly specified. This result agrees with the work of [29] on the comparison among different propensity score methods for estimating the marginal relative risk. Our proposed variance estimation also provides valid type I error and high power. We note that matching can also have smaller bias than other methods, although it may target estimand in a different population that is closer to the treated group. In addition, it has more variability and lower power due to reduced sample size in the matched data and rareness of

the outcomes.

When the propensity score is misspecified, regression adjustment of propensity score with standardization can have more bias. However, the magnitude of the bias is similar to most of the other existing methods evaluated in our simulation studies. Direct adjustment of all the confounders in an outcome regression model with standardization has similar performance as regression adjustment of the propensity score. A potential limitation is that in our simulation setting that mimics the real data example, the number of potential confounders were moderate. Further study is needed to investigate the case when the number of covariates is large enough such that fitting a outcome regression on all covariates is on the edge of having model convergence issues. AIPTW has the worst performance with higher bias across all scenarios. This is likely due to the instability arises from the inversion of the propensity score and rareness of the outcome. When we apply all of the aforementioned methods to a real world application, we see that the estimated effects are similar across different methods. In terms of standard error, regression on propensity score B-spline with standardization has the smallest standard error, indicating potentially increased efficiency in the finite sample setting.

In the setting of rare outcome but common exposure, we suggest the following: as a first step, focus on fitting the propensity score model parametrically to reduce dimensionality; as a second step, apply flexible regression adjustment of propensity score with standardization to control for confounding, rather than traditional propensity score methods such as the augmented IPTW. We also suggest regression on propensity score deciles with standardization as a sensitivity analysis as it fits another nonlinear function of the propensity score and can perform well under the alternative.

Table 4.1: Prevalence (%) of Each Confounder, Relationship between Exposure (ACEI and BB) and Confounders (Propensity Score Model) for Different Simulation Scenarios, and Relationship between Outcome and the Exposure and Confounders (Outcome Model).

		Propensity Score Model (OR)				**Outcome Model (OR)**	
Confounders	Prev %	Observed	Observed+ Interaction	Stronger Propensity	Stronger+ Interaction	Observed	Observed+ Interaction
Heart Disease	2.0	0.5	0.5	0.4	0.4	0.7	0.7
Aspirin	4.4	1.2	1.2	1.3	1.3	0.9	0.9
Ischemic HD	5.3	0.3	0.3	0.2	0.2	1.4	1.4
OptRec Colon Surg	5.6	1.0	1.0	1.0	1.0	1.4	1.4
Inpatient Hosp.	7.9	0.3	0.3	0.2	0.2	1.9	1.9
Allergic Reaction	8.3	0.9	0.9	0.8	0.8	0.6	0.6
NSAIDS	11.8	1.0	1.0	1.0	1.0	5.4	5.4
Diabetes	15.6	4.5	–	9.6	–	2.4	–
Female	51.3	0.6	0.6	0.4	0.4	1.5	1.5
Age(Ref: 18-44)							
45-54	26.6	2.0	–	2.7	–	0.8	–
55-64	29.7	2.1	–	2.9	–	0.5	–
65-99	22.3	1.7	–	2.2	–	0.5	–
Age*Diabetes (Ref: 18-44 and Not Diabetic)							
45-54 & Not Diabetic	22.5		2.0		2.7		0.8
55-64 & Not Diabetic	25.1		2.1		2.9		0.5
65-99 & Not Diabetic	18.8		1.7		2.2		0.5
18-44 & Diabetic	3.3		4.5		9.6		2.4
45-54 & Diabetic	4.1		14.8		42.7		1.2
55-64 & Diabetic	4.6		15.6		45.9		0.7
65-99 & Diabetic	3.5		12.6		34.8		0.7
Exposure							
ACEI	69.1					5.4	5.4

Table 4.2: Bias on log(OR) scale, Type I Error, and Power in Estimating the Marginal OR by Method Ranging the Strength of Confounding and Relationship between Exposure and Outcome, with Correctly Specified Propensity Score Model and Outcome Regression Model.

	Null: log(OR) = 0, OR = 1				Alternative: log(OR) = 1.1, OR = 3			
	Original		Stronger		Original		Stronger	
Methods	Bias (Std Err)	Type I Error	Bias (Std Err)	Type I Error	Bias (Std Err)	Power	Bias (Std Err)	Power
Exposure rate = 69%								
Matching - ATT	-0.012 (0.485)	0.042	-0.014 (0.498)	0.036	-0.001 (0.385)	0.912	-0.021 (0.400)	0.881
Matching - ATE	-0.014		-0.008		-0.005		-0.019	
Augmented IPTW	0.018 (0.433)	0.061	0.039 (0.464)	0.067	0.015 (0.387)	0.871	0.038 (0.423)	0.861
Standardization with Outcome Regression on								
Covariates	0.009 (0.408)	0.049	0.002 (0.407)	0.053	0.005 (0.351)	0.907	-0.000 (0.357)	0.866
PS Linearly	0.011 (0.417)	0.054	-0.001 (0.422)	0.060	0.030 (0.356)	0.909	0.034 (0.367)	0.871
PS Deciles	0.013 (0.403)	0.046	0.011 (0.407)	0.052	0.009 (0.346)	0.908	0.008 (0.354)	0.867
PS B-Splines	0.010 (0.409)	0.049	-0.001 (0.409)	0.051	0.004 (0.351)	0.905	-0.003 (0.357)	0.866
Exposure rate = 20%								
Matching - ATT	-0.000 (1.062)	0.028	0.000 (1.774)	0.026	0.000 (0.754)	0.690	0.000 (1.375)	0.632
Matching - ATE	0.000		0.019		0.021		0.011	
Augmented IPTW	-0.044 (0.952)	0.111	-0.123 (0.710)	0.165	-0.027 (0.357)	0.873	-0.055 (0.427)	0.784
Standardization with Outcome Regression on								
Covariates	-0.006 (0.760)	0.066	-0.008 (0.691)	0.093	-0.002 (0.325)	0.869	0.002 (0.343)	0.807
PS Linearly	-0.001 (0.753)	0.061	-0.012 (0.679)	0.090	-0.018 (0.319)	0.857	-0.041 (0.331)	0.783
PS Deciles	-0.009 (0.784)	0.062	-0.025 (0.713)	0.095	-0.004 (0.325)	0.864	-0.006 (0.347)	0.803
PS B-Splines	0.000 (0.767)	0.065	-0.010 (0.691)	0.095	-0.001 (0.326)	0.865	0.000 (0.347)	0.807

Table 4.3: Bias on log(OR) scale, Type I Error, and Power in Estimating Marginal OR When True Propensity Score Model Has Interactions

	Null: log(OR) = 0, OR = 1				Alternative: log(OR) = 1.1, OR = 3			
	Original		Stronger		Original		Stronger	
Methods	Bias (Std Err)	Type I Error	Bias (Std Err)	Type I Error	Bias (Std Err)	Power	Bias (Std Err)	Power
Exposure rate = 69%								
Matching - ATT	-0.022 (0.479)	0.043	-0.018 (0.471)	0.036	-0.010 (0.381)	0.909	-0.013 (0.378)	0.894
Matching - ATE	-0.012		-0.004		-0.004		-0.010	
Augmented IPTW	0.022 (0.439)	0.061	0.048 (0.451)	0.072	0.017 (0.395)	0.857	0.048 (0.408)	0.880
Standardization with Outcome Regression on								
Covariates	0.008 (0.405)	0.052	0.010 (0.389)	0.054	0.004 (0.350)	0.888	-0.007 (0.336)	0.889
PS Linearly	0.011 (0.414)	0.056	0.006 (0.402)	0.058	0.026 (0.354)	0.889	0.027 (0.345)	0.891
PS Deciles	0.015 (0.401)	0.053	0.010 (0.386)	0.050	0.012 (0.346)	0.890	0.009 (0.333)	0.890
PS B-Splines	0.006 (0.407)	0.054	0.006 (0.391)	0.055	0.004 (0.351)	0.886	-0.005 (0.336)	0.887
Exposure rate = 20%								
Matching - ATT	-0.000 (1.092)	0.030	-0.000 (2.314)	0.014	-0.006 (0.901)	0.706	-0.041 (1.585)	0.541
Matching - ATE	0.018		-0.030		0.015		-0.010	
Augmented IPTW	-0.044 (0.617)	0.120	-0.172 (0.784)	0.183	-0.014 (0.352)	0.873	-0.063 (0.461)	0.718
Standardization with Outcome Regression on								
Covariates	0.003 (0.621)	0.065	-0.015 (0.805)	0.098	0.001 (0.315)	0.875	-0.007 (0.363)	0.766
PS Linearly	0.008 (0.604)	0.065	-0.021 (0.782)	0.091	-0.024 (0.307)	0.858	-0.054 (0.347)	0.718
PS Deciles	-0.004 (0.625)	0.065	-0.034 (0.826)	0.097	0.000 (0.317)	0.866	-0.007 (0.368)	0.753
PS B-Splines	0.007 (0.611)	0.068	-0.022 (0.804)	0.102	0.005 (0.317)	0.873	-0.004 (0.368)	0.756

Table 4.4: Bias on log(OR) scale, Type I Error, and Power in Estimating Marginal OR When True Outcome Model Has Interactions

	Null: log(OR) = 0, OR = 1				Alternative: log(OR) = 1.1, OR = 3			
	Original		Stronger		Original		Stronger	
Methods	Bias (Std Err)	Type I Error	Bias (Std Err)	Type I Error	Bias (Std Err)	Power	Bias (Std Err)	Power
Exposure rate = 69%								
Matching - ATT	-0.021 (0.515)	0.036	-0.008 (0.518)	0.030	-0.018 (0.412)	0.861	-0.016 (0.415)	0.850
Matching - ATE	-0.010		-0.003		-0.008		-0.010	
Augmented IPTW	0.003 (0.463)	0.060	0.022 (0.474)	0.065	0.005 (0.405)	0.860	0.025 (0.423)	0.852
Standardization with Outcome Regression on								
Covariates	0.004 (0.441)	0.050	0.012 (0.431)	0.050	0.002 (0.376)	0.890	-0.003 (0.374)	0.868
PS Linearly	0.001 (0.457)	0.058	0.010 (0.456)	0.064	0.034 (0.385)	0.895	0.049 (0.391)	0.875
PS Deciles	-0.001 (0.436)	0.052	0.011 (0.431)	0.049	0.005 (0.372)	0.888	0.007 (0.372)	0.865
PS B-Splines	0.000 (0.442)	0.053	0.008 (0.434)	0.051	-0.002 (0.376)	0.888	-0.011 (0.375)	0.864
Exposure rate = 20%								
Matching - ATT	-0.000 (1.778)	0.022	-0.000 (2.764)	0.019	0.000 (1.068)	0.640	-0.034 (1.874)	0.505
Matching - ATE	-0.004		-0.031		0.006		-0.020	
Augmented IPTW	-0.045 (1.445)	0.101	-0.137 (1.338)	0.165	-0.022 (0.366)	0.852	-0.066 (0.454)	0.730
Standardization with Outcome Regression on								
Covariates	-0.006 (1.270)	0.064	-0.034 (1.430)	0.098	-0.003 (0.335)	0.862	0.005 (0.374)	0.769
PS Linearly	-0.007 (1.238)	0.063	-0.027 (1.381)	0.098	-0.006 (0.332)	0.853	-0.025 (0.364)	0.751
PS Deciles	-0.009 (1.279)	0.063	-0.044 (1.422)	0.100	-0.002 (0.335)	0.858	-0.008 (0.376)	0.757
PS B-Splines	-0.009 (1.245)	0.062	-0.031 (1.378)	0.103	-0.002 (0.336)	0.858	0.001 (0.375)	0.767

Table 4.5: Estimation and Inference for a Marginal Odds Ratio (ATE) comparing ACEI and BB on angioedema.

Methods	OR (ATE)	Std Err log(OR) scale	Risk of Angioedema ACEI	Risk of Angioedema BB	P value	95% CI
Unadjusted	4.715	0.382	0.157%	0.033%	<0.001	(2.230, 9.967)
Matching (ATT)	6.762	0.536	0.205%	0.030%	<0.001	(2.366, 19.324)
AIPTW	6.477	0.421	0.160%	0.025%	<0.001	(2.841, 14.770)
Standardization						
Covariates	5.677	0.421	0.167%	0.029%	<0.001	(2.490, 12.946)
PS Linearly	5.670	0.420	0.167%	0.029%	<0.001	(2.488, 12.923)
PS Deciles	5.612	0.418	0.166%	0.030%	<0.001	(2.473, 12.735)
PS B-Splines	5.485	0.413	0.165%	0.030%	<0.001	(2.441, 12.326)

Std Err: standard error on the log(OR) scale.

Chapter 5

INFLUENCE AND CORRECTION OF PROVIDER-LEVEL CLUSTERING

5.1 Introduction

There is growing awareness of the potential issue of lack of statistical independence in electronic health record (EHR)-based data specifically. In particular, patients treated by the same provider are more likely to receive similar treatments and therefore respond in a similar manner. In addition, patient characteristics may also vary across providers and if these factors are not adjusted for then differences in patient panels may induce clustering of outcomes. As such, patient outcomes may be correlated within a provider. Ignoring the correlation among patient outcomes within the same provider can lead to a sequence of issues in both the design and analysis of a study. At the design stage, failure to account for the multilevel structure may lead to an underpowered study with an insufficient sample size [74]. When analyzing the data, correlation typically inflates standard errors and can lead to inefficient estimates and anti-conservative inference. Specifically, intra-cluster correlation and analysis that ignores clustering may lead to underestimated standard errors, P-values that are too small, and confidence intervals that are too narrow [27]. For example, in a cross-sectional analysis of patient encounters and referrals during a 1-year period, data from a primary care network of 9 clinics were collected to examine the association between comorbidity and physician referral tendencies. The analysis investigated the need to account for the potential clustering effect of physicians and clinics and found that proper correlated data analysis led to a five-fold increase in the variance of estimated coefficient, indicating a huge impact on inference if one ignores clustering [18].

In this chapter, we consider a simple correction of the standard error estimator discussed

in Chapter 4 to account for potential clustering among patients treated by the same provider. We also evaluate the possible impact of ignoring provider-level clustering, and assess the performance of the correction of standard error through an extension of the realistic simulation study conducted in Chapter 4 by introducing correlation within provider clusters.

5.2 Variance Estimation Under Provider-level Clustering

5.2.1 Provider-level Influence Function Based Variance Estimation

In Chapter 4 we discussed a convenient statistical representation that facilitates estimation of the variance of the marginal, causal odds ratio, by utilizing the influence functions of asymptotically linear estimators. Specifically, the influence function for the log of odds ratio (OR) is $\mathrm{IF}_{\log \mathrm{OR}} = \frac{\mathrm{IF}_1}{\hat{p}_1(1-\hat{p}_1)} - \frac{\mathrm{IF}_0}{\hat{p}_0(1-\hat{p}_0)}$, where IF_1 and IF_0 are the influence functions for the means of risks in the population under treatment and control, $E[Y(1)]$ and $E[Y(0)]$. The variance of an estimate of the log OR can be estimated as a simple scaling of the sample variance of the influence functions, that is,

$$\hat{\sigma}^2(\mathrm{IF}_{\log \mathrm{OR}}) = \frac{1}{n-1}\sum_{i=1}^{n} \mathrm{IF}_i^2,$$

where $i = 1, ...n$ denotes individual patients. Note that such a variance estimator is derived under the assumption that all the n observations are independent.

To account for the potential provider-level clustering, we consider a simple correction of the variance estimator that is similar to a Generalized Estimating Equation (GEE) type sandwich variance estimator with working independence covariance matrix [26]. In particular, the cluster-correlated variance can be estimated as

$$\hat{\sigma}_*^2(\mathrm{IF}_{\log \mathrm{OR}}) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \delta(i,j)\mathrm{IF}_i\mathrm{IF}_j$$

where $\delta(i,j) = 1$ if patients i and j are treated by the same provider and 0 otherwise. By combining terms within a provider cluster we can see that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \delta(i,j)\mathrm{IF}_i\mathrm{IF}_j = \sum_{p=1}^{n_c}\Big(\sum_{i=1}^{m_c} \mathrm{IF}_i^p\Big)^2,$$

where IF_i^p denotes the influence function for patient i treated by provider p, n_c is the total number of clusters, m_c is the size of a cluster, i.e. the number of observations within a cluster. The final standard error is estimated by

$$\mathrm{se}_*(\mathrm{IF}_{\log \mathrm{OR}}) = \hat{\sigma}_*(\mathrm{IF}_{\log \mathrm{OR}})/\sqrt{n} = \sqrt{\frac{1}{n-1}\sum_{p=1}^{n_c}\big(\sum_{i=1}^{m_c}\mathrm{IF}_i^p\big)^2}/\sqrt{n}.$$

We observe that the standard error satisfies

$$\mathrm{se}_*(\mathrm{IF}_{\log \mathrm{OR}}) \approx \sqrt{\frac{1}{n_c-1}\sum_{p=1}^{n_c}\bar{\mathrm{IF}}_p^2}/\sqrt{n_c},$$

where $\bar{\mathrm{IF}}_p = \frac{1}{m_c}\sum_{i=1}^{mc}\mathrm{IF}_i^p$ is the averaged influence function within a provider cluster p of size m_c. Therefore, intuitively, correction for clustering is made by collapsing the influence functions of all patients within the same provider cluster into an averaged *provider-level influence function*, and then calculating the sample variance of these provider-level influence functions.

5.2.2 Multilevel and Non-nested Clustering

EHR-based research may also have a complicated multilevel structure with multiple layers of possible correlations among the outcomes of patients treated by the same physician, from physicians practicing within the same hospital, and from hospitals located within the same geographic region. In addition, patients may have multiple medical complaints that require different providers. Therefore, both multilevel nested clustering and multiple non-nested clustering may exist in healthcare data [13, 62, 63]. As discussed in Chapter 4, the GEE type variance estimator with working independence covariance matrix introduced in [62, 63] can be applied to account for non-nested clustering and multilevel nested clustering. Because the standard error is essentially estimated from the variance of the sample average $\frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}_i$, for a sum of dependent IF's, one can acknowledge multiple sources of correlations introduced by either non-nested clustering or multilevel nested clustering by simply including additional terms corresponding to the correlation between outcomes and associated IF's.

For example, suppose there are two levels of clustering induced by primary care provider and specialty care provider. Then the variance can be estimated by

$$\hat{\sigma}^2_*(\mathrm{IF}_{\log \mathrm{OR}}) = \frac{1}{n-1} \sum_{p=1}^{n_p} \sum_{q=1}^{n_q} \sum_{s=1}^{n_s} \sum_{t=1}^{n_t} \delta(p,q,s,t) \mathrm{IF}_i^{p,s} \mathrm{IF}_j^{q,t},$$

where $\delta(p,q,s,t) = 1$ if $p = q$ or $s = t$, indicating that patients i and j are treated by either the same primary care provider p or the same specialty care provider s, and otherwise $\delta(p,q,s,t) = 0$. By the inclusion-exclusion principle in combinatorics [2],

$$\delta(p,q,s,t) = \mathbb{1}\{p = q\} + \mathbb{1}\{s = t\} - \mathbb{1}\{p = q, s = t\}.$$

Therefore, $\hat{\sigma}^2_*$ can be represented as

$$\hat{\sigma}^2_* = \hat{\sigma}^{2,\mathbb{1}\{p=q\}}_* + \hat{\sigma}^{2,\mathbb{1}\{s=t\}}_* - \hat{\sigma}^{2,\mathbb{1}\{p=q,s=t\}}_*,$$

where $\hat{\sigma}^{2,\mathbb{1}\{p=q\}}_*$ can be obtained from by clustering on primary care provider p, $\hat{\sigma}^{2,\mathbb{1}\{s=t\}}_*$ can be obtained by clustering on specialty care provider s, and $\hat{\sigma}^{2,\mathbb{1}\{p=q,s=t\}}_*$ can be obtained by generating a new cluster-identifying variable (ID) that indicates patients treated by the same primary care provider and same specialty care provider.

The logical representation using the inclusion-exclusion principle is generalizable to more than two levels, and therefore this approach can be extended to more than two levels of non-nested clustering. We can also see that if there are multilevel clusters that are nested, the variance matrix accounting for all clustering levels is equal to the variance matrix accounting for the largest cluster.

5.3 Simulation Study

To evaluate the impact of ignoring provider-level clustering and to investigate the performance of the variance correction, we performed a simulation study following Chapter 4 which mimics real data from the FDA Sentinel Initiative study comparing the effect of angiotensin-converting enzyme Inhibitors (ACEI) and beta blockers (BB) on incidence of angioedema in

the first 30 days [69]. We assign each patient to a provider randomly, and introduce correlation between patients within provider by including a provider-level random intercept that follows a normal distribution with mean zero and variances ranging over a set of values that induce various strength of correlation among observations within provider clusters.

5.3.1 Simulation Setting

We generated a realistic population of $n = 100,000$ patients nested within $n_c = 10,000$ provider clusters, with a fixed cluster size of $m_c = 10$ patients per provider. In our motivating example, there are nine clinically relevant covariates (NSAIDs (Nonsteroidal anti-inflammatory drugs), aspirin, ORAL-CS (optimizing recovery after laparoscopic colon surgery), allergic reaction, diabetes, heart disease, Ischemic HD (heart disease), inpatient hospitalization, and gender) and one categorical variable which is age categorized into four levels.

Specifically, we used the following procedure to generate multilevel data with patients nested within providers. To describe the data structure, let i denote patient and j provider.

1. Generate binary and categorical covariates $\mathbf{Z}_{ij}$ that have the same mean and pairwise covariance as the real data, yielding correlated confounders.

   In total we have 12 confounders (all are binary indicators) in the model including nine binary variables and three indicators for age categories.

2. Generate a binary exposure X (ACEI = 1 and BB = 0) based on a logistic regression on the covariates (the propensity score model, see Table 4.1), using coefficients observed from fitting the real data.

   We include a random intercept a_j shared by patients within the same provider j to introduce correlation among choices of treatments by the same provider. That is, $X_{ij}|\mathbf{Z}_{ij} \sim$ Bernoulli $(p = \text{expit}\{\mathbf{Z}_{ij}\alpha + a_j\})$ where the random intercept $a_j \sim N(0, \sigma_a^2)$.

In addition, we control the fixed intercept such that the exposure prevalence is the same as the real data, which is 69% for ACEI.

3. Generate a pair of binary potential outcomes $(Y_{ij}(1), Y_{ij}(0))$ (angioedema within 30 days under exposure and control for the same subject) based on a logistic regression on the exposure and covariates (the outcome regression model, see Table 4.1), using the coefficients observed from fitting the real data.

   We again include a random intercept b_j to introduce correlation among outcomes of patients treated by the same provider. That is, $Y_{ij}(1)|\mathbf{Z}_{ij} \sim$ Bernoulli $(p = \text{expit}\{\mathbf{Z}_{ij}\beta + \beta_0 + b_j\})$ and $Y_{ij}(0)|\mathbf{Z}_{ij} \sim$ Bernoulli $(p = \text{expit}\{\mathbf{Z}_{ij}\beta + b_j\})$, where the random intercept $b_j \sim N(0, \sigma_b^2)$. We used the simulated pair of potential outcomes to calculate the true ATE and ATT. For methods comparison we used the observed outcome only, which is $Y_{ij} = X_{ij}Y_{ij}(1) + (1 - X_{ij})Y_{ij}(0)$. We hold the event rate in the control group (BB group) the same as the real data for all cases, to generate either a rare outcome with an event rate of 0.03%, or a common outcome with an event rate of 3%. When the outcome event rate in the control group is 0.03%, among $100,000$ subjects there will be about 30 angioedema incidences. When the outcome event rate in the control group is 3%, among $100,000$ subjects there will be about 3000 angioedema incidences.

The similarity among patients within the same provider is often measured by the intra cluster correlation (ICC), a larger value of which indicates stronger correlation. The impact of clustering is determined jointly by the ICC and the size of the clusters. Specifically, when each provider-cluster contains m patients, the ratio of the variance under provider-level clustering to the variance under independence is $1 + (m - 1)\rho$, which is often referred to as the design effect. It is clear from this formula that a small ICC can still substantially influence the validity of inference when the sizes of clusters are large. In our simulation study, to vary the strength of correlation ρ, we set the variances of the provider-level random intercepts σ_a^2, σ_b^2 to be equal to $(0, 1)$, $(1, 1)$, or $(0, 2)$. Note that when σ_a^2 or σ_b^2 is equal to zero, no correlation is induced; when σ_a^2 is nonzero, there will be similar treatment choices

by the same provider; when σ_b^2 is nonzero, patients treated by the same provider will have similar outcomes. With a fixed cluster size of 10, the variance becomes more inflated as the ICC ρ increases. For example, when $\rho = 0.05$, we have that the actual variance of the estimator is 45% larger than the variance under independent data.

We generated the potential outcomes under both the null model ($\log(\text{OR}) = 0$, i.e., no elevated risk of angioedema by treatment of ACEI), and the alternative using coefficient $\log(\text{OR}) = 1.1$ ($\text{OR} = 3$) which represents moderate to strong exposure effect. Under each of the scenarios, we constructed a wale-type testing statistic using the marginal OR estimators and the variance estimator. For the marginal OR estimators we considered the following: (1) "Xadj" = Regression on covariates with standardization; (2) "PS-lin" = regression on main term of the propensity score (linear adjustment) with standardization; (3) "PS-Bspl" = flexible regression of the propensity score using B-spline basis functions with standardization; (4) "PS-str" = regression on propensity score deciles with standardization; (5) "AIPTW" = augmented inverse probability of treatment weighted estimator, with parametric models for exposure and outcome both adjusting for main terms of all covariates, with stabilized propensity score by trimming at 5% tail; and (6) "TMLE" = targeted maximum likelihood estimation. In each scenario assessed we used 2000 simulated datasets. For variance estimator we considered both the estimator that accounts for provider-level clustering, i.e. $\sigma_*^2(\text{IF}_{\log \text{OR}})$, and the one that ignores clustering, i.e. $\sigma^2(\text{IF}_{\log \text{OR}})$, which are detailed in Section 5.2.1.

5.3.2 Results

Figure 5.1 shows the type I error rate of the Wald-type tests. The marginal odds ratio is estimated using methods discussed in Section 5.3.1, and standard error is estimated using the variance estimators discussed in Section 5.2.1. The first row considers rare outcomes and the second row corresponds to non-rare outcomes. The three columns of the panel plot from left to right correspond to the scenarios when variances of the provider-level random intercepts (σ_a^2, σ_b^2) are set to $(0, 1)$, $(1, 1)$, and $(0, 2)$. Interestingly, when the treatment changes from

independent to correlated, indicated by the change of σ_a^2 from 0 to 1 while σ_b^2 is fixed at 1, the ICC increases from 0.02 to 0.03 for rare outcomes, and 0.10 to 0.12 for non-rare outcomes. That is, similar treatment choices by the same provider will introduce additional correlation among patient outcomes even if one fixes the random intercept in the outcomes.

First of all, it is clearly seen that ignoring the correlation within provider clusters will inflate the type I error. The inflated type I error ranges from 0.07 to 0.15. In addition, the degree of inflation increases as the ICC increases, which confirms that the clustering will have more impact when the correlation is stronger. Secondly, the variance estimator that accounts for provider-level clustering can control the type I error at or below the nominated level under ranging strengths of intra cluster correlation, particularly when the outcome is not rare. When the outcome is rare, the correction still provides certain protection against strongly anti-conservative results, although there is certain inflation in the type I error due to more variability from insufficient information in rare outcomes. Third, among the six methods that all estimate the marginal odds ratio and estimate the variance using the influence function with estimated parameters values, AIPTW and TMLE tend to have higher type I error and can be have inflated type I error when the clustering effect is large and the outcomes are not rare, whereas flexible regression on the estimated propensity score (PS-Bspl and PS-str) tend to be more conservative.

Figure 5.2 shows the power of the Wald-type tests. We did not show the scenario when the event rate is sufficiently high which generates non-rare outcomes, because there was no observable power loss among all methods under ranging strength of intra cluster correlation. However, when outcomes are rare, as shown in Figure 5.2, the corrected tests has lower power than the anti-converative tests that ignores clustering, although the corrected inference still have reasonable power to detect a significant treatment effect.

It is also of notice that when $\sigma_a^2 = 1$ and $\sigma_b^2 = 1$, the ICC is larger than when $\sigma_a^2 = 0$ and $\sigma_b^2 = 1$. This implies that the correlation induced among treatment mechanism will have an impact downstream on the correlation among outcomes. This make sense because when the provider tend to provide similar treatments to patients, the outcome will be more correlated

than if the treatments are completely independent.

5.4 Application

We apply statistical methods discussed in Section 5.3 to the Back Pain Outcomes using Longitudinal data (BOLD) study, which is a prospective study that enrolled 5,239 patients that are 65 or older with a new episode of back pain, and with a primary interest in the cost-effectiveness of early diagnostic imaging among the elderly [Jarvik et. al. (2015)]. Specifically, whether older adults with back pain should undergo early imaging has been controversial. On one hand, there is a higher prevalence of serious underlying conditions such as cancer which necessitate early diagnosis. On the other hand, there is also a high prevalence of incidental findings which may lead to unnecessary interventions that increase costs without corresponding benefits. Thus, it is critical to understand whether early imaging improves patients health outcomes. Both patient-reported outcomes and EHRs were collected, and patients are recruited through their primary care provider [Jarvik et. al. (2012)]. Therefore, there is potential correlation among patient outcomes within the same provider. The primary outcome was the Roland-Morris Disability Questionnaire (RMDQ), a measure of physical limitations due to back pain. In this section, we take 30% improvement of RMDQ over a year as our outcome of interest, which is a binary outcome that indicates a clinically meaningful benefit. We are looking for the association between early radiographic imaging and clinical benefit in one-year RMDQ. A summary of patients' baseline characteristics comparing treatment groups is shown in Table 5.1, which presents a few baseline covariates that are highly imbalanced, such as study site, age, baseline diagnosis, and so on. Therefore there is need to account for the systematic differences among patient baseline characteristics. We apply causal inference methods discussed in Chapter 4 to estimate the effect of early imaging while controling for potential confounding. There are 2006 primary care providers with an average provider-level cluster size of 2.3. We estimated an intra cluster correlation (ICC) of 0.05 which indicates the potential need to consider accounting for clustering. We estimate a marginal odds ratio comparing the rate of improvement in RMDQ

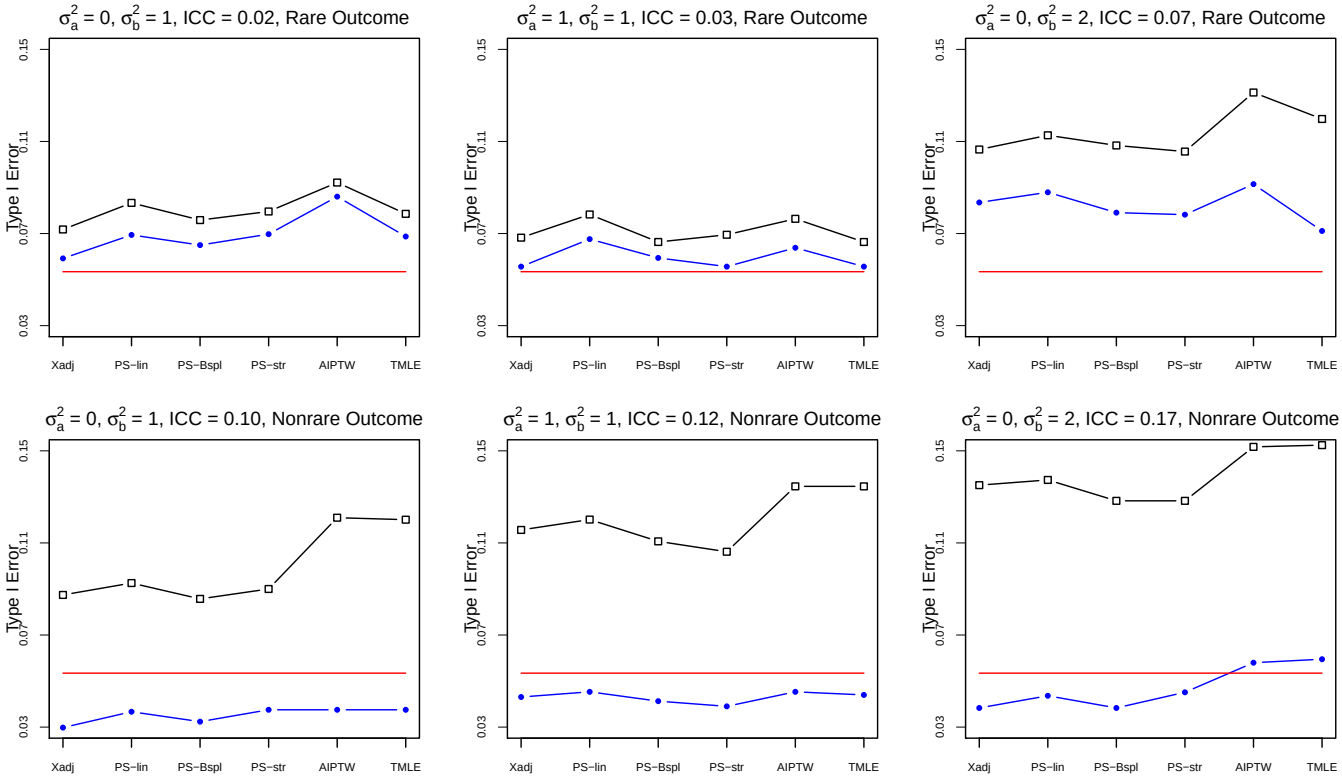

Figure 5.1: Type I error rate of the Wald-type test using estimated variances accounting for or ignoring provider-level clustering, under different estimating methods, and under ranging variance of the random intercepts σ_a^2 and σ_b^2 which induce different strength of correlation implied by the intra cluster correlation (ICC). Colored lines correspond to nominated α-level of 0.05 (——) ; ignoring clustering (▫—▫—▫) ; correct for clustering (●—●—●) . The methods for estimating the marginal odds ratio are: "Xadj" = Regression on covariates with standardization; "PS-lin" = regression on main term of the propensity score (linear adjustment) with standardization; "PS-Bspl" = flexible regression of the propensity score using B-spline basis functions with standardization; "PS-str" = regression on propensity score deciles with standardization; "AIPTW" = augmented inverse probability of treatment weighted estimator, with parametric models for exposure and outcome both adjusting for main terms of all covariates, with stabilized propensity score by trimming at 5% tail; "TMLE" = targeted maximum likelihood estimation.

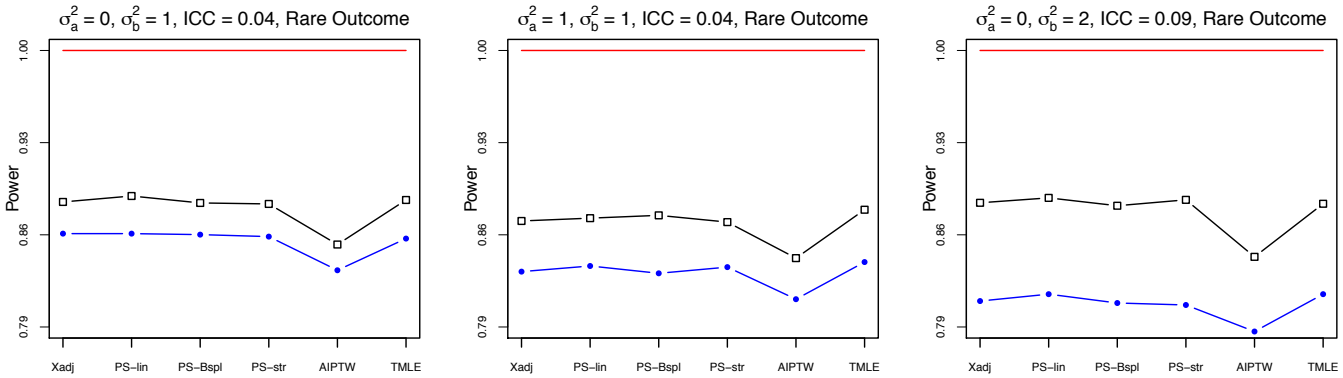

Figure 5.2: Power of the Wald-type test using estimated variances accounting for or ignoring provider-level clustering, under different estimating methods, and under ranging variance of the random intercepts σ_a^2 and σ_b^2 which induce different strength of correlation implied by the intra cluster correlation (ICC). Colored lines correspond to nominated α-level of 0.05 (——) ; ignoring clustering (▫—▫—▫) ; correct for clustering (●—●—●) . The methods for estimating the marginal odds ratio are: "Xadj" = Regression on covariates with standardization; "PS-lin" = regression on main term of the propensity score (linear adjustment) with standardization; "PS-Bspl" = flexible regression of the propensity score using B-spline basis functions with standardization; "PS-str" = regression on propensity score deciles with standardization; "AIPTW" = augmented inverse probability of treatment weighted estimator, with parametric models for exposure and outcome both adjusting for main terms of all covariates, with stabilized propensity score by trimming at 5% tail; "TMLE" = targeted maximum likelihood estimation.

between patients with and without early imaging. In particular, our inference will consider clustering by primary care provider.

Table 5.2 presents the estimated marginal log odds ratios along with the sandwich standard error that is robust to provider-level clustering. We also present the estimated standard error ignoring correlation within primary care provider. We can see that, when patients may belong to the same provider group, the naive standard error ignoring provider-level clustering can be 8% smaller, leading to an anti-conservative interpretation. In contrast, accounting for correlation within provider cluster can provide a valid result that is robust to clustering.

5.5 Discussion

Although there is increasing consideration about clustering in group-randomized trials, there has been less discussion about provider-level clustering in EHR-based observational research. In this chapter, we consider a simple correction of the variance estimator of marginal estimates that has a causal interpretation in observational studies. The corrected variance estimation accounts for correlation among patient outcomes within the same provider by collapsing the influence functions of patients within the same provider cluster into an averaged influence function, followed by estimating the sample variance of provider-level influence functions. In the realistic simulation study, we learned that similar treatment choices by the same provider will introduce additional correlation among patient outcomes even if one fixes the random intercept in the outcomes. We also observe that when patients treated by the same provider tend to have similar treatments and similar outcomes, ignoring such similarity will result in an anti-conservative statistical inference with inflated type I error and larger than expected power. The impact of clustering increases as the intra cluster correlation increases. In contrast, our simple correction can control the type I error with reasonable statistical power to detect a significant treatment effect.

Table 5.1: Baseline Characteristics: Early Radiographic Imaging vs. No Early Radiographic Imaging Patients

Variable	Level	N	Combined (N = 4709)	No-Early Radiograph (N = 3429)	Early Radiograph (N = 1280)	P-val
Site, No. (%)		4709				<0.001
	Henry Ford		841 (18%)	558 (16%)	283 (22%)	
	Kaiser		3077 (65%)	2206 (64%)	871 (68%)	
	Harvard Vanguard		791 (17%)	665 (19%)	126 (10%)	
Female, No. (%)		4709				0.935
	Yes		3056 (65%)	2227 (65%)	829 (65%)	
Race, No. (%)		4660				0.422
	Black/African American		725 (15%)	545 (16%)	180 (14%)	
	Asian		187 (4%)	132 (4%)	55 (4%)	
	Caucasian		3430 (73%)	2484 (72%)	946 (74%)	
	Other/Multiple		318 (7%)	233 (7%)	85 (7%)	
	Missing		49 (1%)	35 (1%)	14 (1%)	
Hispanic, No. (%)		4687				1
	Yes		287 (6%)	209 (6%)	78 (6%)	
	Missing		22 (0%)	16 (0%)	6 (0%)	
Age, mean +/- sd		4709	73.7 +/- 6.8	73.4 +/- 6.7	74.4 +/- 7.0	<0.001
Education, No. (%)		4700				0.004
	<High school		278 (6%)	190 (6%)	88 (7%)	
	High school, trade school, or some college		2473 (53%)	1762 (51%)	711 (56%)	
	College graduate		1147 (24%)	874 (25%)	273 (21%)	
	Graduate degree		802 (17%)	596 (17%)	206 (16%)	
	Missing		9 (0%)	7 (0%)	2 (0%)	

Variable	Level	N	Combined (N = 4709)	No-Early Radiograph (N = 3429)	Early Radiograph (N = 1280)	P-val
Smoking status, No. (%)		4700				0.903
	Never smoked		2593 (55%)	1890 (55%)	703 (55%)	
	Quit>1 year ago		1821 (39%)	1328 (39%)	493 (39%)	
	Current smoker		286 (6%)	205 (6%)	81 (6%)	
	Missing		9 (0%)	6 (0%)	3 (0%)	
Married/Living with partner, No. (%)		4697				0.257
	Yes		2849 (61%)	2093 (61%)	756 (59%)	
	Missing		12 (0%)	7 (0%)	5 (0%)	
Baseline Diagnosis, No. (%)		4709				<0.001
	Non-Specific Back Pain Only		3189 (68%)	2323 (68%)	866 (68%)	
	Back and Leg Pain		1023 (22%)	716 (21%)	307 (24%)	
	Lumbar Stenosis		233 (5%)	201 (6%)	32 (2%)	
	Other		264 (6%)	189 (6%)	75 (6%)	
Prior Imaging, No. (%)		4709	0.1 +/- 0.3	0.1 +/- 0.3	0.0 +/- 0.2	<0.001
Pain Duration, No. (%)		4708				<0.001
	<1 month		1605 (34%)	1141 (33%)	464 (36%)	
	1-3 month		928 (20%)	619 (18%)	309 (24%)	
	3-6 month		320 (7%)	207 (6%)	113 (9%)	
	6-12 month		281 (6%)	193 (6%)	88 (7%)	
	1-5 years		680 (14%)	534 (16%)	146 (11%)	
	> 5 years		894 (19%)	734 (21%)	160 (12%)	
	Missing		1 (0%)	1 (0%)	0 (0%)	
Quan Comorbidity, No (%)		4595				0.01
	0-1 comorbidity		638 (14%)	440 (13%)	198 (15%)	
	2 comorbidities		1702 (36%)	1277 (37%)	425 (33%)	
	> 2 comorbidities		2255 (48%)	1632 (48%)	623 (49%)	
	Missing		114 (2%)	80 (2%)	34 (3%)	

Variable	Level	N	Combined (N = 4709)	No-Early Radiograph (N = 3429)	Early Radiograph (N = 1280)	P-val
Confidence pain will improve in 3 months, No. (%)		4701				<0.001
	0 (not at all confident)		956 (20%)	748 (22%)	208 (16%)	
	1-4		669 (14%)	468 (14%)	201 (16%)	
	5		748 (16%)	545 (16%)	203 (16%)	
	6-9		1301 (28%)	890 (26%)	411 (32%)	
	10 (extremely confident)		1027 (22%)	773 (23%)	254 (20%)	
	Missing		8 (0%)	5 (0%)	3 (0%)	
Baseline back pain NRS (0-10), mean +/- sd		4709	5.1 +/- 2.8	4.9 +/- 2.8	5.5 +/- 2.7	<0.001
Baseline leg pain NRS (0-10), mean +/- sd		4708	3.5 +/- 3.3	3.4 +/- 3.3	3.7 +/- 3.4	0.003
Baseline ≧1 fall past 3 weeks, No. (%)		4707				<0.001
	Yes		350 (7%)	228 (7%)	122 (10%)	
	Missing		2 (0%)	1 (0%)	1 (0%)	
Baseline RMDQ mean +/- sd		4709	9.6 +/- 6.4	9.2 +/- 6.5	10.9 +/- 6.0	<0.001
Baseline BPI, mean +/- sd		4709	3.3 +/- 2.5	3.2 +/- 2.5	3.8 +/- 2.4	<0.001
Baseline EQ5D Index, mean +/- sd		4701	0.8 +/- 0.2	0.8 +/- 0.2	0.7 +/- 0.2	<0.001
Baseline PHQ4, mean +/- sd		4695	1.6 +/- 2.5	1.5 +/- 2.4	1.8 +/- 2.6	<0.001
Total RVUs prior year, mean +/- sd		4701	37.2 +/- 86.7	36.9 +/- 86.0	38.2 +/- 88.7	0.63

T-test and Chi-Square test are used to obtain p-values; sd=standard deviation; No.=number; Prior imaging: Defined as spine imaging at the healthcare site in 365 days prior to index visit.; RMDQ: Roland-Morris Disability Questionnaire. Range is 0 (no pain-related limitations) - 24 (maximum pain-related limitations); BPI: Brief Pain Inventory Interference. Range is 0 (no pain interference)-10 (maximum pain interference); EQ-5D Index:EuroQuol Group Index score. Range is 0 (death)-1 (perfect health); NRS: Numeric Rating Scale. Range is 0 (no pain)-10 (pain as bad as you can imagine); RVU: Relative value unit

Table 5.2: Marginal log Odds Ratio estimated by Causal Inference Methods

Method	**ATE**	**robust se**	**robust CI**	**naive se**	**naive CI**
Xadj	-0.078	0.085	(-0.244,0.088)	0.081	(-0.237,0.082)
PS-lin	-0.090	0.099	(-0.285,0.105)	0.092	(-0.270,0.090)
PS-Bspl	-0.084	0.099	(-0.278,0.111)	0.091	(-0.262,0.095)
PS-str	-0.100	0.100	(-0.296,0.096)	0.093	(-0.282,0.081)
AIPTW	-0.070	0.085	(-0.236,0.096)	0.081	(-0.230,0.089)
TMLE	-0.067	0.084	(-0.233,0.098)	0.081	(-0.226,0.091)

Chapter 6

CONCLUDING REMARKS AND FUTURE WORK

6.1 Conclusion

In the United States, the use of electronic health records is now incentivized due to the 2009 enactment of the Health Information Technology for Economic and Clinical Health (HITECH) Act. Large scale EHR data opens new opportunities for research to improve patient care and the health of the public. Current national research efforts include linking EHR to conduct pharmacosurveillance (e.g., FDA Sentinel) and assembling large clinical populations for comparative effectiveness research (e.g., PCORnet). However, EHR data are not collected for research purposes. Facilitating the use of administrative EHR data to address comparative effectiveness and safety questions creates numerous challenges. This dissertation focuses on the following three particular challenges. First, there is potential for variation in both the medical coding practice and the choice of specific healthcare procedures, due to differences in patient populations and/or financial incentives within care delivery networks. Second, pharmacosurveillance research using EHR data entails the need for robust causal inference to estimate drug adverse effects controlling for confounding. In this setting, we often encounter an extremely rare adverse outcome and a large number of confounders. Conventional covariate adjustment methods may have convergence issues when the outcome is rare and flexible nonparametric regression may be challenging due to the high-dimensionality of the covariates. Third, a unique and often omitted feature of EHR data is its hierarchical structure. Specifically, EHR-based research may have a multilevel structure with multiple layers of correlations among the outcomes of patients treated by the same physician, physicians practicing within the same hospital, and hospitals located within the same geographic region. Failure to account for potential correlation among patients within provider clusters

may lead to both poorly designed study and invalid statistical inference.

In this dissertation, we explore methods for multivariate inference and surveillance to facilitate the secondary use of large-scale EHR data for research purpose. In Chapter 2, we detail statistical testing and estimation procedures to compare the endorsement of CPT codes, in order to identify potential data quality issue, and to investigate differential patterns in healthcare utilization. With the unique hierarchical structure in terms of both thematically grouped medical codes and provider-level clustering, we consider penalized regression methods unifying estimation and inference with hierarchical shrinkage to leverage such structure and stabilize estimates for rare procedures.

In Chapter 3 and 4, we consider the particular setting of rare outcomes with many confounders that is frequently encountered in postmarketing surveillance. We first review existing literatures on methodology and theory for regression adjustment using the propensity score as a covariate, which provides both sufficient control of confounding and dimension reduction. We detail previous studies on semiparametric efficient estimation and explore recent advances in econometric literature on the semiparametric efficient estimation with generated covariates. We finally provide a formal statistical representation of the three-step estimator using the influence function, and extend existing asymptotic analysis to assess the asymptotic properties of the three-step estimation with a generated covariate which is the propensity score.

The influence function representation of the three-step estimator allows one to directly estimate the variance of the marginal causal effect in a target population. In Chapter 5, we consider potential correlation among patients within a provider cluster and extend the standard error such that it is robust to provider-level clustering. The robust standard error can control the type I error and can be generalized to multilevel and potentially nonnested clustering.

6.2 Future Work

This dissertation aims at solving the unique challenges in using population-scale EHR data for research purpose and contributes to facilitating the learning of healthcare systems, healthcare utilization, as well as the monitoring of safety of new medical products. However, the multivariate inference and causal inference methods may be limited in interpretation. Particularly, the multivariate inference methods may require further investigation of whether the statistical finding corresponds to actual differences in patient care, or whether coding variation through use of alternative codes may explain differences in observed endorsement rates. In addition, the three-step estimator may detect a statistically significant result that is in fact not clinically meaningful. Therefore, it is critical to carefully interpret the results and translate statistical findings back into scientific knowledge.

Another future direction is to consider the settings of longitudinal and survival data. Although this dissertation focuses on EHR-based data extracted by collapsing in time, it is scientifically important to acknowledge that the EHR data evolves over time and to draw inferences from a complex, real-world healthcare system incorporating time. The longitudinal feature of EHR data allows one to analyze change over time, to predict risks flexibly, and to actively monitor drug safety. Therefore a future direction is to extend existing methods to longitudinal settings. Moreover, in large population-based datasets, the proportional hazard assumption is frequently violated. In addition, a population-level HR with causal interpretation cannot be simply adapted from the conditional HR which is entangled in time and covariates. An alternative is to estimate treatment-specific population-level survival curves as two-dimensional nonparametric functions of time and the propensity score. Then any causal contrast can be done by definition to measure a select treatment effect. Therefore this novel statistical inference for survival data via estimation of treatment-specific flexible survival curves is another future direction.

Forth, the high dimensionality of patient features recorded in EHR data presents both challenges and opportunities that may be further studied. In particular, the high-dimensional

covariates such as demographics, diagnoses, and medical procedures are the common causes of both the outcome and the treatment mechanism. Such dual roles can be summarized in a prognostic score which improves precision, and a propensity score which reduces bias. The prognostic score and the propensity score are both balancing scores that potentially removes confounding but with different advantages. Building a regularized and prognostically balanced propensity score may lead to a sufficient statistic that improves over the original balancing scores. In addition, study of valid statistical inference after variable selection is also critical. Therefore, it is of interest to further improve the three-step estimation procedure as well as to study the asymptotic properties when there are multiple generated covariates or potential variable selection in the first step of propensity score estimation.

Lastly, one critical issue with nonparametric regression is the selection of smoothing parameters. A well-known procedure commonly used in practice is cross-validation. Recently, with the advances of influence function-based statistical representation, there has been some proposal on targeted smoothing which selects the smoothing parameters that targets the bias-variance tradeoff for the parameter of interest rather than intermediate summaries [38]. Further investigation on the consistency and efficiency of nonparametric estimators with targeted smoothing could provide theoretical foundation and practical recommendation that facilitates the application of flexible modeling strategies.

BIBLIOGRAPHY

[1] Ralph B Agostino D. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med, 17(19):2265–2281, 1998.

[2] Reginald BJT Allenby and Alan Slomson. How to count: An introduction to combinatorics. CRC Press, 2011.

[3] P. C. Austin. The performance of different propensity score methods for estimating marginal odds ratios. Statistics in medicine, 26(16):3078–3094, 2007.

[4] P. C. Austin, P. Grootendorst, S. L. T. Normand, and G. M. Anderson. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a monte carlo study. Statistics in medicine, 26(4):754–768, 2007.

[5] Peter C Austin. The performance of different propensity score methods for estimating marginal odds ratios. Statistics in medicine, 26(16):3078–3094, 2007.

[6] Peter C Austin. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. American journal of epidemiology, 172(9):1092–1097, 2010.

[7] Peter C Austin. The performance of different propensity score methods for estimating marginal hazard ratios. Statistics in medicine, 32(16):2837–2849, 2013.

[8] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973, 2005.

[9] S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. Genetic Epidemiology, 35(7):606–619, 2011.

[10] R.E. Behrman, J.S. Benner, J.S. Brown, M. McClellan, J. Woodcock, and R. Platt. Developing the sentinel system — a national resource for evidence development. N Engl J Med, 2011.

[11] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. Journal of Econometrics, 186(2):345–366, 2015.

[12] P.N. Bentley, A.G. Wilson, M.E. Derwin, R. Scodellaro, and R.E. Jackson. Reliability of assigning correct current procedural terminology-4 e/m codes. Ann Emerg Med., 40:269–74, 2002.

[13] Rebecca A Betensky, James A Talcott, and Jane C Weeks. Binary data with two, non-nested sources of clustering: an analysis of physician recommendations for early prostate cancer treatment. Biostatistics, 1(2):219–230, 2000.

[14] C.A. Boneau. The effects of violations of assumptions underlying the t test. Psychological bulletin, page 49, 1960.

[15] P. Bühlmann. Statistical significance in high-dimensional linear models. Bernoulli, 19:1212–1242, 2013.

[16] Shelley B Bull. Regression models for multiple outcomes in large epidemiologic studies. Statistics in medicine, 17(19):2179–2197, 1998.

[17] J. Chapman and J. Whittaker. Analysis of multiple SNPss in a candidate gene or region. Genetic Epidemiology, 32(6):560–566, 2008.

[18] Frederick M Chen, George E Fryer, and Thomas E Norris. Effects of comorbidity and clustering upon referrals in primary care. The Journal of the American Board of Family Practice, 18(6):449–452, 2005.

[19] Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. Econometrica, 71(5):1591–1608, 2003.

[20] Guang Cheng, Lan Zhou, Jianhua Z Huang, et al. Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data. Bernoulli, 20(1):141–163, 2014.

[21] P. E. Cheng. Nonparametric estimation of mean functionals with data missing at random. Journal of the American Statistical Association, 89(425):81–87, 1994.

[22] William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A, pages 417–446, 1973.

[23] Carl De Boor. B (asic)-spline basics. Technical report, DTIC Document, 1986.

[24] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. A practical guide to splines, volume 27. New York: Springer-Verlag, 1978.

[25] Paul Dierckx. Curve and surface fitting with splines. Oxford University Press, 1995.

[26] P.J. Diggle, P.and Liang K-Y. Heagerty, and S.L. Zeger. Analysis of Longitudinal Data. Oxford University Press, New York, 2002.

[27] Allan Donner. Some aspects of the design and analysis of cluster randomization trials. Journal of the Royal Statistical Society: Series C (Applied Statistics), 47(1):95–113, 1998.

[28] Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. Annals of Statistics, 27(5):1491–1518, 1999.

[29] Jessica M Franklin, Wesley Eddings, Peter C Austin, Elizabeth A Stuart, and Sebastian Schneeweiss. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. Statistics in Medicine, 2017.

[30] M. Frydenberg. Marginalization and collapsibility in graphical interaction models. The Annals of Statistics, pages 790–805, 1990.

[31] Lorentz G.G. Approximations of Function. New York: Chelsea, 1986.

[32] S Greenland. Introduction to regression modelling. Chapter 21. In: Rothman KJ, Greenland S, Lash TL (eds). Lippincott Williams & Wilkins, 2008.

[33] Sander Greenland. Collapsibility. In International Encyclopedia of Statistical Science, pages 267–270. Springer, 2011.

[34] Sander Greenland, James M Robins, and Judea Pearl. Confounding and collapsibility in causal inference. Statistical Science, pages 29–46, 1999.

[35] Susan Gruber and Mark J van der Laan. tmle: An r package for targeted maximum likelihood estimation. 2011.

[36] Roee Gutman and Donald B Rubin. Estimation of causal effects of binary treatments in unconfounded studies. Statistics in medicine, 34(26):3381–3398, 2015.

[37] Erinn M Hade and Bo Lu. Bias associated with using the estimated propensity score as a regression covariate. Statistics in medicine, 33(1):74–87, 2014.

[38] Jenny Häggström and Xavier de Luna. Targeted smoothing parameter selection for estimating average causal effects. Computational Statistics, 29(6):1727–1748, 2014.

[39] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica, pages 315–331, 1998.

[40] Jinyong Hahn, Zhipeng Liao, and Geert Ridder. Nonparametric two-step sieve m estimation and inference. 2016.

[41] Jinyong Hahn and Geert Ridder. Asymptotic variance of semiparametric estimators with generated regressors. Econometrica, 81(1):315–340, 2013.

[42] Ben B Hansen. The prognostic analogue of the propensity score. Biometrika, 95(2):481–488, 2008.

[43] Wolfgang Härdle. Applied nonparametric regression. Number 19. Cambridge university press, 1990.

[44] Wolfgang Härdle, Hua Liang, and Jiti Gao. Partially Linear Models. New York: Springer., 2000.

[45] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological), pages 757–796, 1993.

[46] Xuming He and Peide Shi. Monotone b-spline smoothing. Journal of the American statistical Association, 93(442):643–650, 1998.

[47] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12:55–67, 1970.

[48] J. Holt, A. Warsy, and P. Wright. Medical decision making: guide to improved CPT coding. South Med J, 103(4):316–322, 2010.

[49] Donald R Hoover, John A Rice, Colin O Wu, and Li-Ping Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika, 85(4):809–822, 1998.

[50] J.G. et. al. Jarvik. Study protocol: the back pain outcomes using longitudinal data (BOLD) registry. BMC Musculoskelet Disord., 13, 2012.

[51] M.S. King, M.S. Lipsky, and L. Sharp. Expert agreement in Current Procedural Terminology evaluation and management coding. Arch Intern Med., 162(4):316–320, 2002.

[52] M.S. King, L. Sharp, and M.S. Lipsky. Accuracy of CPT evaluation and management coding by family physicians. J Am Board Fam Pract., 14(462):184–192, 2001.

[53] Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. The Annals of Statistics, pages 1548–1562, 1987.

[54] S. Lee, M. J. Emond, M. J. Bashed, K. C. Barnes, M. J. Rieder, D. A. Nickerson, NHLBI GO Exome Sequencing Project - ESP Lung Project Team, D. C. Christiani, M. M. Wurzel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet, 91(2):224–237, 2012.

[55] Oliver Linton and Jens Perch Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika, pages 93–100, 1995.

[56] Roderick Little and Hyonggin An. Robust likelihood-based analysis of multivariate data with missing values. Statistica Sinica, pages 949–968, 2004.

[57] Berlin JA Localio AR, Margolis DJ. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. J Clinical Epi, 60:874–82, 2007.

[58] B.E. Madsen and S.R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. PLOS Genetics, 5(2):e1000384, 2009.

[59] A. Maity, Y. Ma, and R. J. Carroll. Efficient estimation of population-level summaries in general semiparametric regression models. Journal of the American Statistical Association, 102(477):123–139, 2007.

[60] Enno Mammen, Christoph Rothe, and Melanie Schienle. Semiparametric estimation with generated covariates. Econometric Theory, 32(5):1140–1177, 2016.

[61] Enno Mammen, Christoph Rothe, Melanie Schienle, et al. Nonparametric regression with nonparametrically generated covariates. The Annals of Statistics, 40(2):1132–1170, 2012.

[62] Diana L Miglioretti and Patrick J Heagerty. Marginal modeling of multilevel binary data with time-varying covariates. Biostatistics, 5(3):381–398, 2004.

[63] Diana L Miglioretti and Patrick J Heagerty. Marginal modeling of nonnested multilevel data using standard software. American Journal of Epidemiology, 165(4):453–463, 2007.

[64] R. G. Miller Jr. Beyond ANOVA: basics of applied statistics. CRC Press., 1997.

[65] S. Morgenthaler and W.G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutat Res, 615(1-2):28–56, 2007.

[66] A.P. Morris and E. Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiology, 34(2):188–193, 2010.

[67] U. U. Müller. Estimating linear functionals in nonlinear regression with responses missing at random. Journal of the American Statistical Association, 37(5A):2245–2277, 2009.

[68] Jessica A Myers and Thomas A Louis. Comparing treatments via the propensity score: stratification or modeling? Health Services and Outcomes Research Methodology, 12(1):29–43, 2012.

[69] Jennifer C. Nelson, Denise Boudreau, Robert Wellman, Onchee Yu, Andrea J. Cook, and et. al. Improving sequential safety surveillance planning methods for routine assements that use regression adjustment or weighting to control confounding. https://www.sentinelsystem.org/sentinel/methods/routine-prospective-safety-surveillance-new-drugs-vaccines-and-other-biologic/, 2016. Mini-Sentinel Methods Report.

[70] Whitney K Newey. Series estimation of regression functionals. Econometric Theory, 10(01):1–28, 1994.

[71] W.K. Newey. Kernel estimation of partial means and a general variance estimator. Econometric Theory, 10(2):1–21, 1994.

[72] W.K. Newey. The Asymptotic Variance of Semiparametric Estimators. Econometrica, 62(6):1349–1382, 1994.

[73] W.K. Newey. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics, 79(1):99–135, 1997.

[74] Sharon-Lise T Normand and Kelly H Zou. Sample size considerations in observational health care quality studies. Statistics in medicine, 21(3):331–345, 2002.

[75] W. Pan. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genetic Epidemiology, 33(6):487–507, 2009.

[76] Frank J Potter. A study of procedures to identify and trim extreme sampling weights. In Proceedings of the section on survey research methods, American Statistical Association, pages 225–230, 1990.

[77] Frank J Potter. The effect of weight trimming on nonlinear survey estimates. In Proceedings of the American Statistical Association, Section on Survey Research Methods, volume 758763, 1993.

[78] J. Przyborowski and H. Wilenski. Homogeneity of results in testing samples from Poisson series. Biometrika, 31, 1940.

[79] Y. Qi, D.E. Weeks, H.K. Tiwari, N. Yi, K. Zhang, G. Gao, W. Lin, Lou X., Chen W., and Liu N. Rare-variant kernel machine test for longitudinal data from population and family samples. Hum Hered, 80(3):126–138, 2015.

[80] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling, 7(9):1393–1512, 1986.

[81] James Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. Journal of chronic diseases, 40:139S–161S, 1987.

[82] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. Epidemiology, pages 550–560, 2000.

[83] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427):846–866, 1994.

[84] James M Robins and Naisyin Wang. Inference for imputation estimators. Biometrika, 87(1):113–124, 2000.

[85] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 2010.

[86] M.D. Robinson and G.K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics, 9:321–332, 2008.

[87] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

[88] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American statistical Association, 79(387):516–524, 1984.

[89] Jean Claude Roujeau and Robert S Stern. Severe adverse cutaneous reactions to drugs. New England Journal of Medicine, 331(19):1272–1285, 1994.

[90] Donald B Rubin. On principles for modeling propensity scores in medical research. Pharmacoepidemiology and drug safety, 13(12):855–857, 2004.

[91] David Ruppert, Matt P Wand, and Raymond J Carroll. Semiparametric regression. Number 12. Cambridge university press, 2003.

[92] G.D. Ruxton. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology., 17:688–690, 2006.

[93] Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychological methods, 13(4):279, 2008.

[94] Baiju R Shah, Andreas Laupacis, Janet E Hux, and Peter C Austin. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. Journal of clinical epidemiology, 58(6):550–559, 2005.

[95] X. Shi, H. Pashova, and P. J. Heagerty. Comparing healthcare utilization patterns via global differences in the endorsement of current procedural terminology codes. In press in Annals of Applied Statistics, 2016.

[96] Jonathan M Snowden, Sherri Rose, and Kathleen M Mortimer. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. American Journal of Epidemiology, 173(7):731–738, 2011.

[97] Kelly G Stratton, Andrea J Cook, Lisa A Jackson, and Jennifer C Nelson. Simulation study comparing exposure matching with regression adjustment in an observational safety setting with group sequential monitoring. Statistics in medicine, 34(7):1117–1133, 2015.

[98] Dag Tjøstheim and Bjørn H Auestad. Nonparametric identification of nonlinear time series: projections. Journal of the American Statistical Association, 89(428):1398–1409, 1994.

[99] Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. The International Journal of Biostatistics, (1), 2006.

[100] Aad W Van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 2000.

[101] Stijn Vansteelandt and Rhian M Daniel. On regression adjustment for the propensity score. Statistics in medicine, 33(23):4053–4072, 2014.

[102] Stijn Vansteelandt and Rhian M Daniel. On regression adjustment for the propensity score. Statistics in medicine, 33(23):4053–4072, 2014.

[103] Fei Wan and Nandita Mitra. An evaluation of bias in propensity score-adjusted non-linear regression models. Statistical methods in medical research, pages 1–17, 2016.

[104] Li Wang, Xiang Liu, Hua Liang, and Raymond J Carroll. Estimation and variable selection for generalized additive partial linear models. Annals of statistics, 39(4):1827–1851, 2011.

[105] Edward J Wegman and Ian W Wright. Splines in statistics. Journal of the American Statistical Association, 78(382):351–365, 1983.

[106] H. Wu, C. Wang, and Z. Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics, 14:232–243, 2012.

[107] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet, 89(1):82–93, 2011.

[108] D.W. Zimmerman and B.D. Zumbo. Rank transformations and the power of the Student t test and Welch t'test for non-normal populations with unequal variances. Canadian Journal of Experimental Psychology., 47:523–539, 1993.

Appendix A

APPENDIX FOR CHAPTER 2

A.1 Comprehensive Discussion on Code-wise Two-sample Testing Options

We provide detailed review of testing strategies that are candidates for the evaluation of variation in code endorsement rates across cohorts. Our interest in CPT code-wise inference requires selection of a testing strategy that can be valid for both common and rare codes, and under potential overdispersion. We review standard asymptotic likelihood ratio tests (LRTs) and conditional exact tests (ETs) for count and binary data with a goal of characterizing the applied options.

A.2 Count Outcome Asymptotic and Exact Tests: Simple Poisson Model for Rates

For count data, a natural model is a Poisson distribution. Specifically, the counts for code assignment are assumed to be characterized by a rate parameter, λ_s^c, with $Y_{si}^c \sim \text{Poisson}(\lambda_s^c)$, $i = 1, \ldots, n_s$.

In this situation, we may wish to test whether code endorsement varies by cohort, i.e., H_0: $\lambda_0^c = \lambda_1^c$. In the EMR setting, the number of patients is usually quite large (thousands or greater) and large sample approximations should be valid. Therefore, if the Poisson assumption was valid, then the LRT using observations from each patient could provide inference regarding coding rates across cohorts. However, rare codes can lead to low expected cell counts, which can lead to a poor χ^2 approximation for the null distribution of the LRT. In this case, as an exact alternative, we can collapse patient-level information within a cohort by computing the total counts $Y_s^c = \sum_{i=1}^{n_s} Y_{si}^c \sim \text{Poisson}(n_s\lambda_s^c)$, and apply the Poisson ET which does not depend on large sample approximations [78]. The Poisson ET computes the p-value

based on the fact that, under the null, $Y_1^c \mid Y_0^c + Y_1^c \sim \text{Binomial}(n = Y_0^c + Y_1^c,\ p = \frac{n_1}{n_0+n_1})$.

A.3 Count Outcome Asymptotic and Exact Tests: Negative Binomial Model for Overdispersion

In cohorts where there are both relatively healthy and extremely ill patients, there will be overdispersion in the CPT utilization counts, and the simple Poisson mean-variance relationship will not hold. In this situation, use of negative binomial distribution provides one model-based generalization of the Poisson assumption. The negative binomial model contains a rate parameter, and an additional parameter that characterizes overdispersion: $Y_{si}^c \sim$ negative binomial$(p = p_s^c = \frac{\lambda_s^c}{\lambda_s^c + \frac{1}{\phi^c}}, r = 1/\phi^c)$, $i = 1, \ldots, n_s$, where ϕ^c is the code-specific overdispersion parameter which is shared by patients across cohorts. The mean and variance are parameterized as $\text{E}[Y_{si}^c] = \lambda_s^c$ and $\text{Var}[Y_{si}^c] = \lambda_s^c(1 + \lambda_s^c\phi^c)$. Note that when $\phi^c = 0$, the model reduces to the Poisson model.

The sum of negative binomial variables with the same p is still negative binomial, and since the patient-level count data for code c within a cohort s are i.i.d. with the same p_s^c, the cohort-level data summary counts are also negative binomial: $Y_s^c = \sum_{i=1}^{n_s} Y_{si}^c \sim$ negative binomial$(p' = \frac{n_s\lambda_s^c}{n_s\lambda_s^c + \frac{n_s}{\phi^c}}, r' = \frac{n_s}{\phi^c})$. Note that the probability $p' = p$, while the mean and variance are $\text{E}[Y_s^c] = n_s\lambda_s^c$ and $\text{Var}[Y_s^c] = n_s\lambda_s^c(1 + \lambda_s^c\phi^c)$.

We take parallel asymptotic and exact testing strategies as with the Poisson model. When the LRT χ^2 approximation can be assumed to be valid, we use the negative binomial LRT with patient-level data Y_{si}^c. Note that the overdispersion parameter will inflate the standard errors, making the test results more conservative than the Poisson LRT. When the expected cell counts are too small, we can instead rely on the negative binomial ET to calculate the conditional probability using the cohort-level data Y_s^c [85]. For a given code, we wish to test whether the CPT code assignment varies by cohort, i.e., H_0: $\lambda_0^c = \lambda_1^c \equiv \lambda^c$ without the requirement of asymptotic methods. Under the null, we have $p_0^c = p_1^c = \frac{\lambda^c}{\lambda^c + \frac{1}{\phi^c}}$ and $Y_s^c \sim$ negative binomial$(p' = \frac{\lambda^c}{\lambda^c + \frac{1}{\phi^c}}, r' = \frac{n_s}{\phi^c})$. Thus $Y_0^c + Y_1^c \sim$ negative binomial$(p'' =$

$\frac{\lambda^c}{\lambda^c+\frac{1}{\phi^c}}, r'' = \frac{n_0+n_1}{\phi^c}$). We compute the p-value via the conditional probability

$$\Pr(Y_0^c = y_0^c, Y_1^c = y_1^c \mid Y_0^c + Y_1^c = y_0^c + y_1^c) =$$
$$\binom{y_0^c + \frac{n_0}{\phi^c} - 1}{y_0^c} \cdot \binom{y_1^c + \frac{n_1}{\phi^c} - 1}{y_1^c} \Big/ \binom{y_0^c + y_1^c + \frac{n_0+n_0}{\phi^c} - 1}{y_0^c + y_1^c}.$$

Although the conditional probability eliminates the common rate parameter, λ^c, p-value computation still require the common overdispersion parameter ϕ^c. We can either vary the value of the nuisance parameter over a range of plausible values, or we can estimate ϕ^c based on the patient-level data Y_{si}^c using maximum likelihood. Specifically, we use negative binomial regression methods that allow separate group rates (e.g., do not impose null for λ_0^c and λ_1^c) and estimate the common overdispersion parameter. Because the test statistic has an exact distribution, with large sample sizes we expect substituting a consistent estimator of the nuisance parameter to yield similar performance as if the true nuisance parameter were known, and the sample size of EMR data is often sufficiently large. However, if the sample size is truly insufficient, estimation of the overdispersion parameter with small sample sizes has been studied intensively for differential expression detection in RNA-seq data [86, 106], which can serve as alternative methods. In our simulation studies we evaluate the performance of the negative binomial ET with a plug in dispersion parameter.

The negative binomial distribution is a natural model for overdispersed count data that characterized unobserved heterogeneity through use of a continuous latent variable. A key advantage of the negative binomial is a generalized variance function which may allow valid testing of rate ratios. However, in certain situations the population under study may be comprised of a mixture of discrete subgroups such that a continuous latent variable model does not capture the true heterogeneity. When discrete heterogeneity is present, latent class or mixture models may describe the population distribution for count data, but often will require additional parameters that specify the mixture proportions and the subgroup count distributions. Therefore, the use of mixture models for testing will typically lead to a model that under the null will constrain multiple parameters which leads to a primary test statistic that has multiple degrees of freedom rather than a single focused test statistic for rate ratios.

Power may be poor unless mixture model alternatives are carefully chosen such as fixing the mixture probabilities, or constraining feature of the component mixtures. For example, zero-inflated Poisson models are a special case of a discrete mixture, but have one component of the mixture specified as a point mass at zero. We focus our evaluation on the use of standard overdispersed count data models and their associated tests, but recognize that additional exploration of testing options using more complex mixture models is a direction of important future work. However, we do evaluate the performance of simple Poisson and negative binomial tests when data may in fact be generated from a more complex mixture mechanism such as a zero-inflated Poisson.

A.4 Count Outcome Asymptotic and Exact Tests: Semiparametric Two-sample t-test

While use of the negative binomial model allows a partial decoupling of the mean and the variance, it may not provide valid inference when the true data generating mechanism is not adequately characterized by a simple overdispersed count model. However, with large sample sizes and any underlying distribution, we can use the two-sample t-test as a semi-parametric method for testing. Use of the t-test for inference with CPT codes requires understanding the performance of the method for both common and rare codes, and below we focus on three central issues.

First, with a large overall sample and equal groups sizes, the t-test is known to be robust to both inequality of variance and non-normality, and maintains a valid test size due to the central limit theorem [14]. Second, with comparison groups that have unequal variances, Welch's t-test would be preferred as it correctly estimates the standard deviation in the denominator of the t statistic. In addition, Welch's t-test has been shown to perform better than the Student's t-test (assuming equal variances) when the data are normally distributed [92, 108].

Finally, while the Welch's t-test naturally accommodates unequal variances, it does not necessarily eliminate the effect of non-normality. Although EMR data may provide large

sample sizes, when the outcome rate is rare, the sample mean of discrete counts may not be approximately normally distributed. In particular, the effect of non-normality lies in the higher order moments (skewness and kurtosis) that arise in the expansion of the mean and variance of the t statistic. It has been shown that the kurtosis has little effect on the asymptotic distribution of the t statistic, and has no effect when the sample sizes are equal. The skewness does affect the asymptotic mean and variance of the t statistic. Thus, comparing the sample means for two populations with different shapes requires caution and perhaps necessitates alternative testing strategies [64]. Ultimately, t-tests provide simple asymptotic inference, but the performance in terms of size and power for EMR data should be investigated to guide applied statistical practice for both rare and common outcomes.

A.5 Any/None Outcome Asymptotic and Exact Tests

Use of CPT counts and associated rates of endorsement will often be the appropriate strategy for answering scientific questions about variation in utilization. However, for certain codes such as recommended annual screening measures or vaccinations, it may be desirable to simply analyze the potential count data as a derived binary outcome since the clinical significance is indicated by any endorsement of the code.

Statistically the methods for testing a binary outcome are well characterized. An any/none outcome indicates whether a patient was ever assigned a code during his or her visits over the year. These outcomes can be modeled as $Z^c_{si} \sim \text{Bernoulli}(p^c_s)$ for a patient, and $Z^c_s = \sum_{i=1}^{n_s} Z^c_{si} \sim \text{Binomial}(n_s, p^c_s)$ for a cohort. Our goal is to test whether the probability of assigning a CPT code varies by cohort, that is, H_0: $p^c_0 = p^c_1$. When the requirements of the χ^2 approximation are met for the LRT, we use the Binomial LRT. When the expected cell counts are too small, we can use the conditional ET assuming Binomial model, which is the well-known Fisher's ET for a two-by-two table constructed using cohort level data. Under the null, the conditional distribution of the cell counts is hypergeometric, with conditional

probability

$$\Pr(Z_0^c = z_0^c, Z_1^c = z_1^c \mid Z_0^c + Z_1^c = z_0^c + z_1^c) = \binom{n_0}{z_0^c} \cdot \binom{n_1}{z_1^c} \Big/ \binom{n_0 + n_1}{z_0^c + z_1^c}.$$

In summary, key practical issues include: whether to adopt asymptotic tests or exact tests; whether to consider a count or indicator outcome; and whether or how to account for overdispersion in CPT counts.

A.6 Proof of Lemma 2.1

Proof. Consider the estimating equation of ridge regression for the Poisson family with a log link

$$\frac{1}{n}\mathbf{X}^T(\boldsymbol{Y} - e^{\log(\boldsymbol{t})+\mathbf{X}\boldsymbol{\beta}}) + \boldsymbol{\Lambda}\boldsymbol{\beta} \triangleq \boldsymbol{U}_n(\boldsymbol{\beta}) + \boldsymbol{\Lambda}\boldsymbol{\beta} = \mathbf{0}, \tag{A.1}$$

where we define

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{X}^T(\boldsymbol{Y} - e^{\log(\boldsymbol{t})+\mathbf{X}\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i - e^{\log(t_i)+\mathbf{X}_i^T\boldsymbol{\beta}})$$

is the estimating function of the generalized linear model

$$\log[E(\boldsymbol{Y})] = \log(\boldsymbol{t}) + \mathbf{X}\boldsymbol{\beta}. \tag{A.2}$$

We first state some classical results for the generalized linear model. For the random variables $(\boldsymbol{X}, Y, t)$, we define the function

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{X}(Y - e^{\log(t)+\boldsymbol{X}^T\boldsymbol{\beta}}).$$

Since $\boldsymbol{\beta}^*$ is the population coefficient satisfying (A.2), we have $\mathrm{E}[\boldsymbol{U}(\boldsymbol{\beta}^*)] = 0$. For notational convenience, we let $\boldsymbol{\Omega}(\boldsymbol{\beta}^*) = \mathrm{Var}[\boldsymbol{U}(\boldsymbol{\beta}^*)]$. By the Central Limit Theorem, $\sqrt{n}\boldsymbol{U}_n(\boldsymbol{\beta}^*) \xrightarrow{d} N(0, \boldsymbol{\Omega}(\boldsymbol{\beta}^*))$. We will use these results in the following proof.

We now prove Lemma 2.1. By second order Taylor expansion on (A.1), we obtain

$$0 = [\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Lambda}](\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*) + \boldsymbol{U}_n(\boldsymbol{\beta}^*) + \boldsymbol{\Lambda}\boldsymbol{\beta}^* + o_p(1), \tag{A.3}$$

where $\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) = \mathbf{X}^T \text{diag}(e^{\log(\boldsymbol{t})+\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda})\mathbf{X}/n$ is obtained by taking derivative of $\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_\lambda)$ with respect to $\hat{\boldsymbol{\beta}}_\lambda$. By rearranging the terms in (A.3) and multiplying both sides of the equation by $\sqrt{n}$, we obtain

$$\sqrt{n}\big\{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Lambda}]\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{\beta}^*\big\} = -\sqrt{n}\boldsymbol{U}_n(\boldsymbol{\beta}^*) + o_p(1).$$

Equivalently, we have

$$\sqrt{n}[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Lambda}]\big\{\hat{\boldsymbol{\beta}}_\lambda - [\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Lambda}]^{-1}\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{\beta}^*\big\} \xrightarrow{d} N(0, \boldsymbol{\Omega}(\boldsymbol{\beta}^*)), \tag{A.4}$$

where we use the fact that $\sqrt{n}\boldsymbol{U}_n(\boldsymbol{\beta}^*) \xrightarrow{d} N(0, \boldsymbol{\Omega}(\boldsymbol{\beta}^*))$.

For a given $\boldsymbol{\Lambda}$, let $\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}$ be the population ridge coefficient satisfying $E[\boldsymbol{U}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}) + \boldsymbol{\Lambda}\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}] = 0$. By continuous mapping theorem, we have $\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) \xrightarrow{p} \boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})$. We use the notation $\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})$ as a general representation of the limit of $\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)$, which includes the case when λ may grow with n. However, if λ is fixed and bounded, then in the limit $\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}) = \boldsymbol{H}(\boldsymbol{\beta}^*)$. Thus, by Slutsky's theorem and (A.4), we have

$$\sqrt{n}\Big(\hat{\boldsymbol{\beta}}_\lambda - [\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}) + \boldsymbol{\Lambda}]^{-1}\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})\boldsymbol{\beta}^*\Big) \xrightarrow{d} N(0, [\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}) + \boldsymbol{\Lambda}]^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}^*)[\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}) + \boldsymbol{\Lambda}]^{-1}).$$

Note that $\hat{\boldsymbol{\beta}}_\lambda$ is not an unbiased estimator of $\boldsymbol{\beta}^*$. We propose the following de-bias estimator for estimating $\boldsymbol{\beta}^*$:

$$\hat{\boldsymbol{\beta}}_{\text{debias}} = \big\{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Lambda}]^{-1}\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)\big\}^{-1}\hat{\boldsymbol{\beta}}_\lambda. \tag{A.5}$$

By (A.4) and (A.5),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{debias}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0, [\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})]^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}^*)[\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})]^{-1}).$$

□

A.7 Comprehensive Review of Simulation Results Comparing Group-wise Association Tests

We provide a review of relevant results in previous research comparing group-wise association tests. Generalizing these results to our context implies that burden tests may have increased

power to detect association at the block level when the direction of code-specific effects are similar. In contrast, when code-specific effects may differ in direction, SKAT has been shown to be an effective testing strategy [107]. When provider-level clustering is present, the L-KM method controls the type I error and increases power compared to competing methods [79].

A.7.0.0.1 Independent Data [107] compared the type I error and power of SKAT and burden tests for binary outcomes. In their simulation settings, genetic variants have a variety of frequencies: in the context of CPT code, this corresponds to data with both rare and common codes within a given block. Among the different types of burden tests considered, the counting-based burden test corresponds to our procedure in Section 2.2.2.1. The sum test by [9] also shares the same spirit with our burden test.

Both SKAT and burden tests are able to control the type I error at $\alpha = 0.05$ level. However, they can be conservative when the sample size is small. In terms of power, [107] considered two different cases: (1) all the causal variants have positive effects and (2) 50% - 80% causal variants have positive effects and the rest are negative. For genetic applications the magnitude of the effect is assumed to be determined by the frequency of the variant: rarer variants have larger absolute effects (in either direction). When effects have different directions, SKAT has the highest power while the burden test suffers from low power. Not surprisingly, when all of the variants have effects that are in the same direction, the counting-based burden test has the highest power.

In practice with CPT codes, the choice between the burden test and SKAT for block-level inference should be guided by any available biomedical knowledge on the direction of code effects. Specifically, if within a block certain procedures are heavily used in one cohort, and some are heavily used in the other cohort, e.g., code substitution, then SKAT would be preferred.

A.7.0.0.2 Correlated Data [79] considered longitudinal outcome and compared the longitudinal kernel-machine regression (L-KM) method to other treatments of longitudinal

outcome such as taking the average within cluster and treating the average outcome as independent. The L-KM controls the type I error at numerous significance level, and has the highest power as it uses full information in the data.

A.8 Comprehensive Plots of Type I Error and power

We provide additional supporting plots (see Figures A.1-A.8) that show the type I error and power of all tests with equal/unequal sample sizes using generated data of independent observations or under provider-level clustering. Figures A.1-A.5, and A.7-A.8 have three columns, which correspond to Poisson, Negative Binomial, and Zero-inflated Poisson data respectively. Each row in these figures corresponds to a testing method.

In particular, Fig A.1-A.2 show the type I error using independent observations with equal/unequal sample sizes, whereas Fig A.3-A.4 show the type I error using correlated data that are generated using a a mean-preserving random variable that follows a Normal/Gamma distribution. Fig A.5 demonstrates coverage of confidence interval from ridge regression, and Fig A.6 shows the type I error of the dynamic test. Fig A.7-A.8 display the power using independent observations with equal/unequal sample sizes.

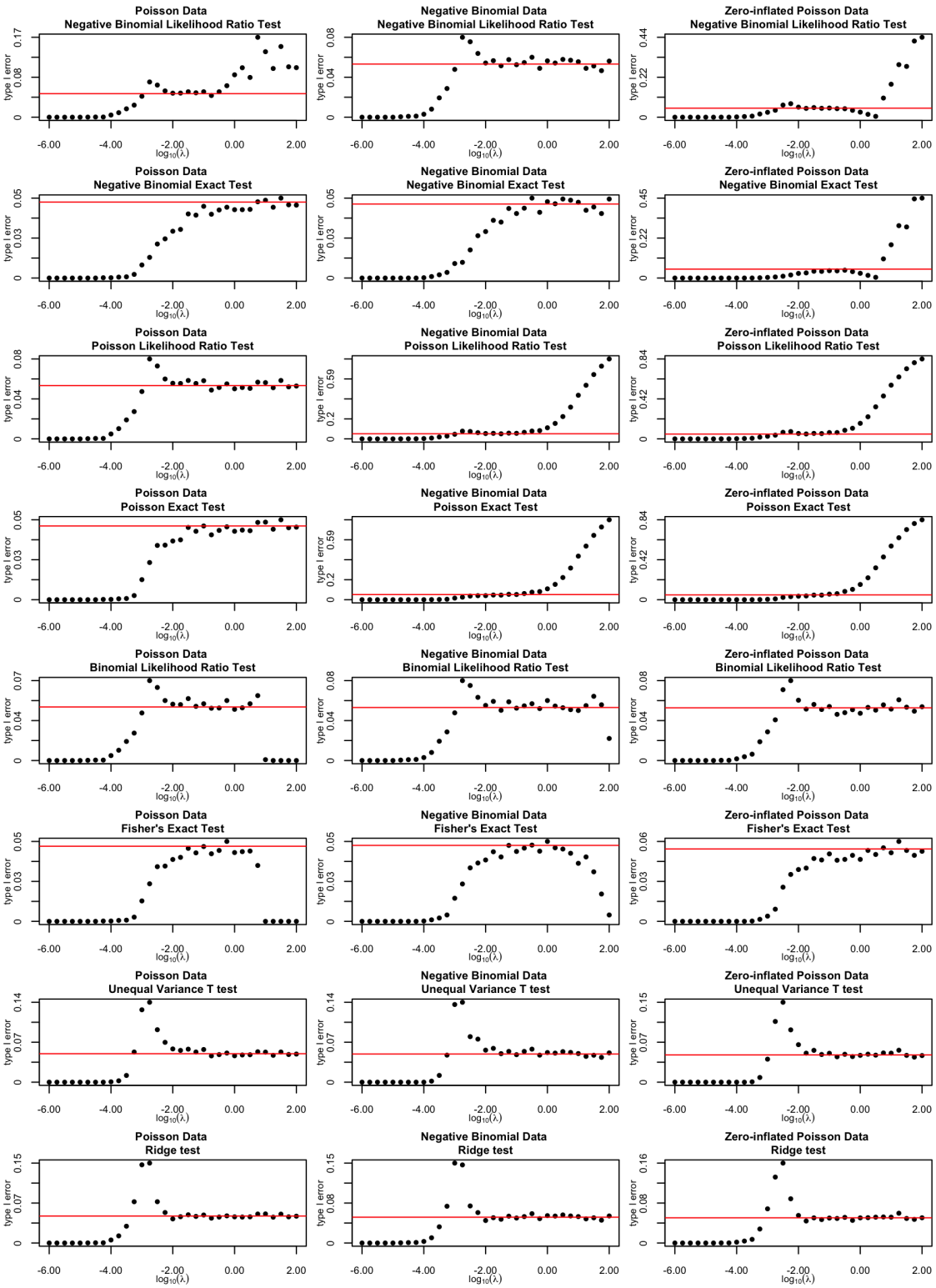

Figure A.1: Type I error rates of all CPT code-specific tests with **unequal** sample sizes ($n_0 = 1000$, $n_1 = 3000$) using Poisson data, Negative Binomial data, or Zero-inflated Poisson data, each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale.

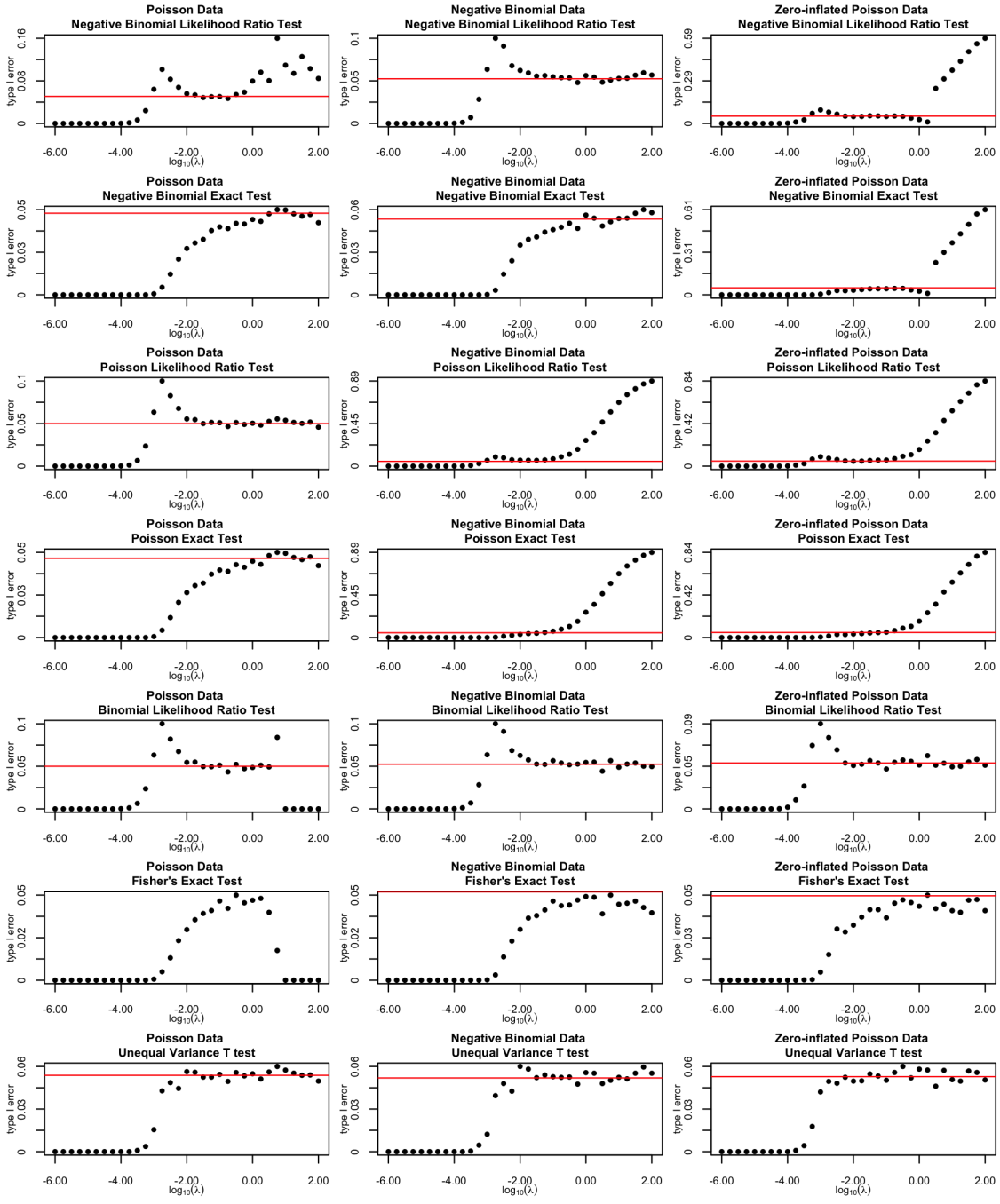

Figure A.2: Type I error rates of all CPT code-specific tests using Poisson data, Negative Binomial data, or Zero-inflated Poisson data with **equal** sample sizes ($n_0 = 1000, n_1 = 1000$), each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale.

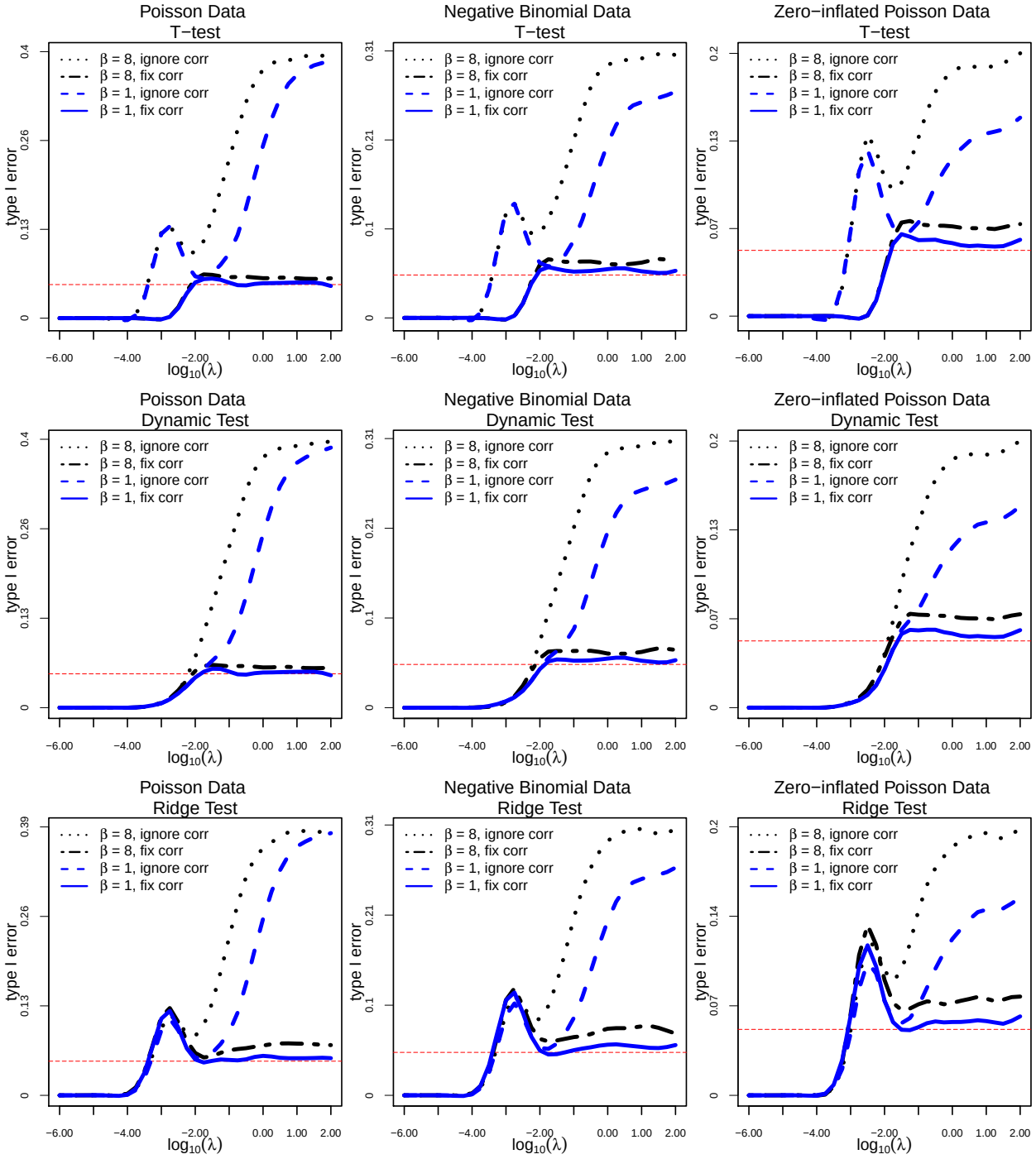

Figure A.3: Type I error rates of t-test, dynamic test, and ridge test with provider-level clustering using Poisson data, Negative Binomial data, or Zero-inflated Poisson data. To introduce association, a mean-preserving random variable $\gamma_p^\beta \sim \text{Gamma}(\text{shape} = \frac{1}{\beta}, \text{scale} = \beta)$ with $\text{E}[\gamma_p^\beta] = 1$ and $\text{Var}[\gamma_p^\beta] = \beta$ is shared by patients treated by provider p. The cohorts have unequal sample sizes ($n_0 = 1,000$, $n_1 = 3,000$), each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale.

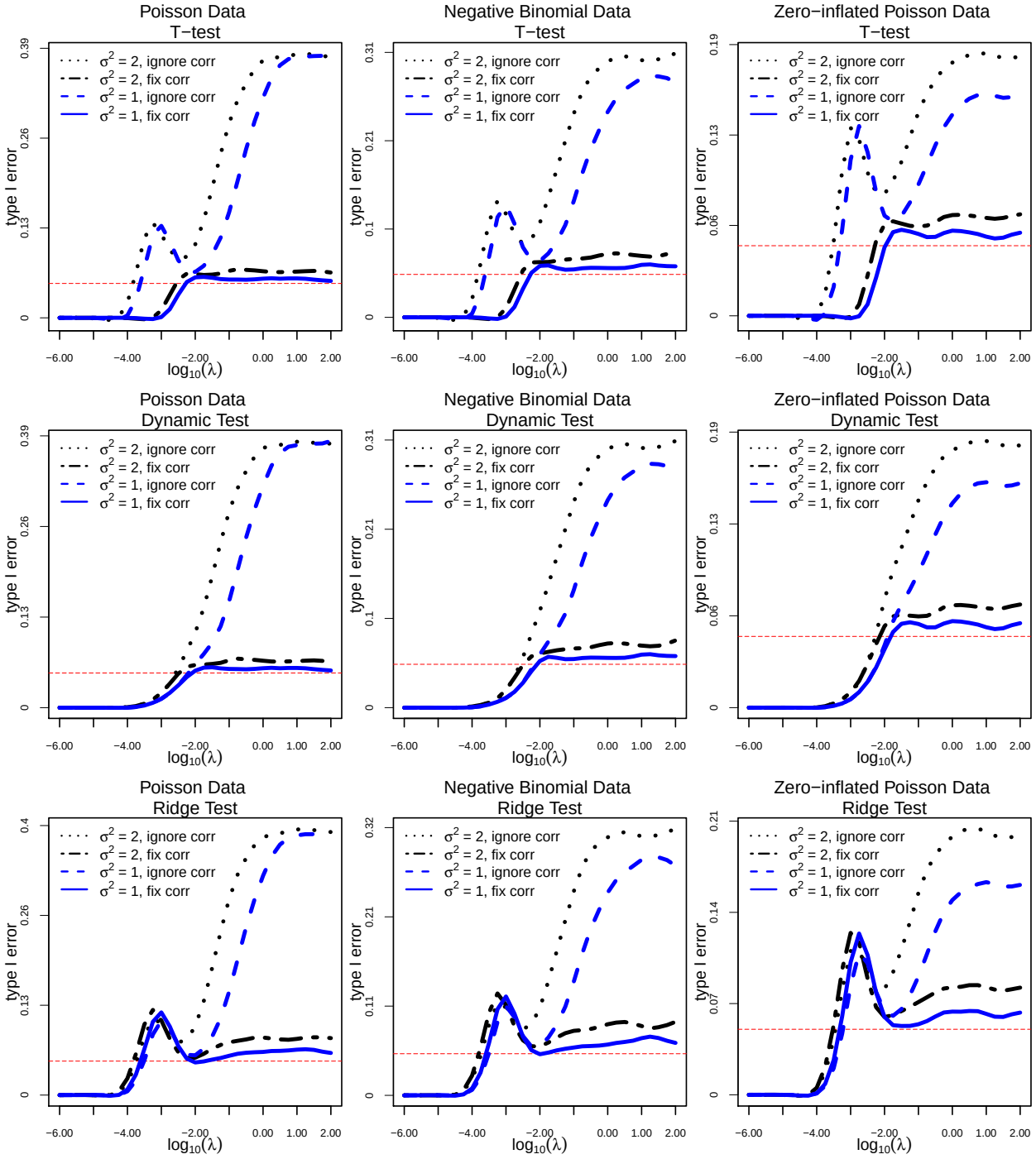

Figure A.4: Type I error rates of t-test, dynamic test, and ridge test with provider-level clustering using Poisson data, Negative Binomial data, or Zero-inflated Poisson data. To introduce association, a mean-preserving random variable $\gamma_p^\sigma \sim \text{Normal}(0, \sigma^2)$ with $\text{E}[\gamma_p^\sigma] = 0$ and $\text{Var}[\gamma_p^\sigma] = \sigma^2$ is shared by patients treated by provider p. The cohorts have unequal sample sizes ($n_0 = 1,000$, $n_1 = 3,000$), each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale.

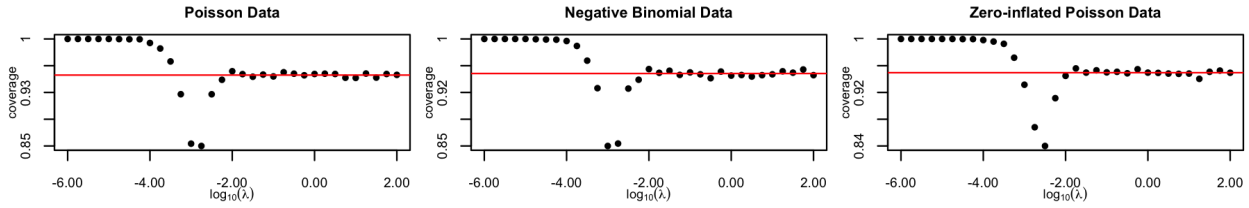

Figure A.5: Coverage of confidence interval for ridge Poisson regression using Poisson data, Negative Binomial data, or Zero-inflated Poisson data with **unequal** sample sizes ($n_0 = 1000$, $n_1 = 3000$), each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale.

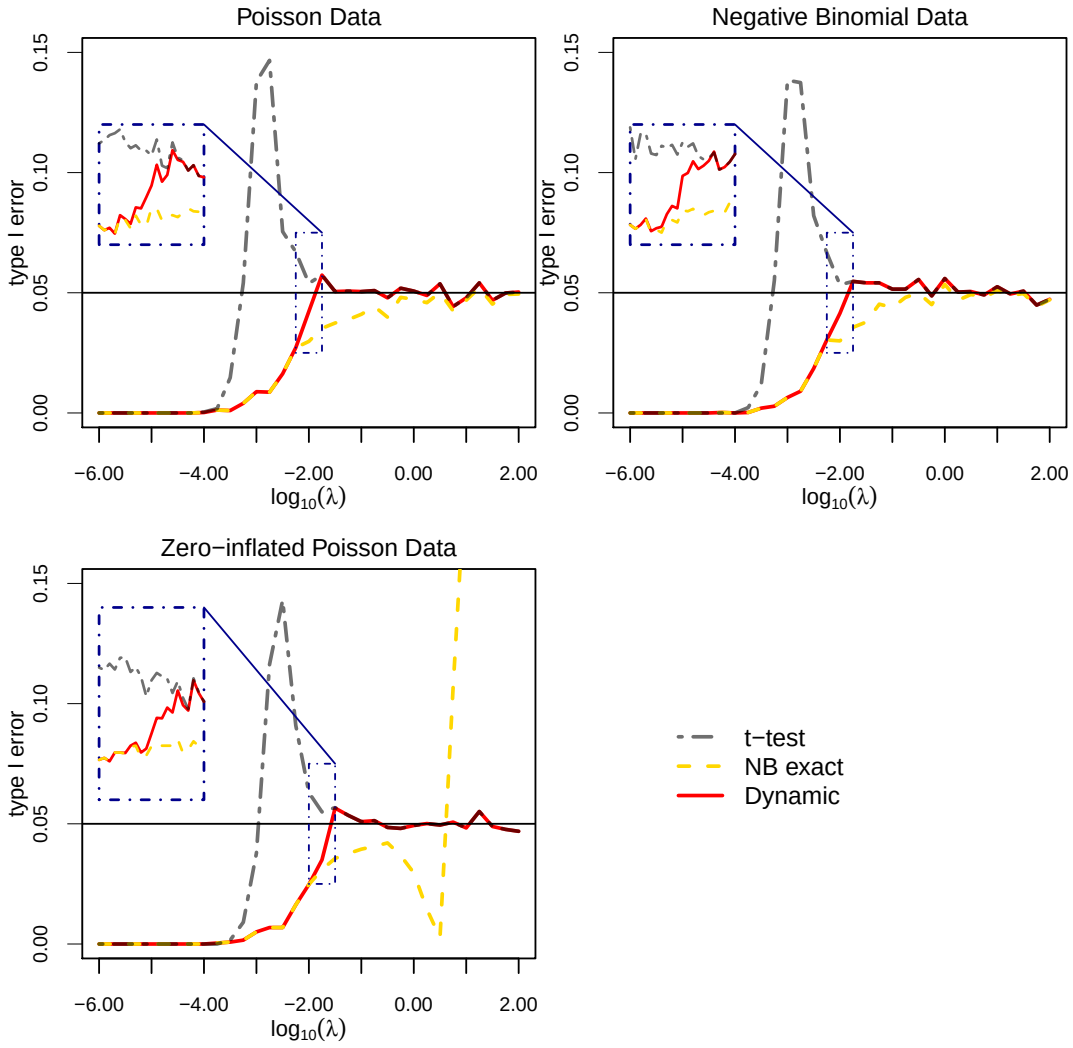

Figure A.6: Type I error rates of the dynamic test which is a mixture of the Negative Binomial exact test and the t-test, with unequal sample sizes ($n_0 = 1000$, $n_1 = 3000$) using Poisson data, Negative Binomial data, or Zero-inflated Poisson data, each with a group mean of λ ranging from 10^{-6} to 10^2, plotted on $\log_{10}$ scale. Colored lines correspond to Negative Binomial ET (— —); t-test (— —); Dynamic test (——).

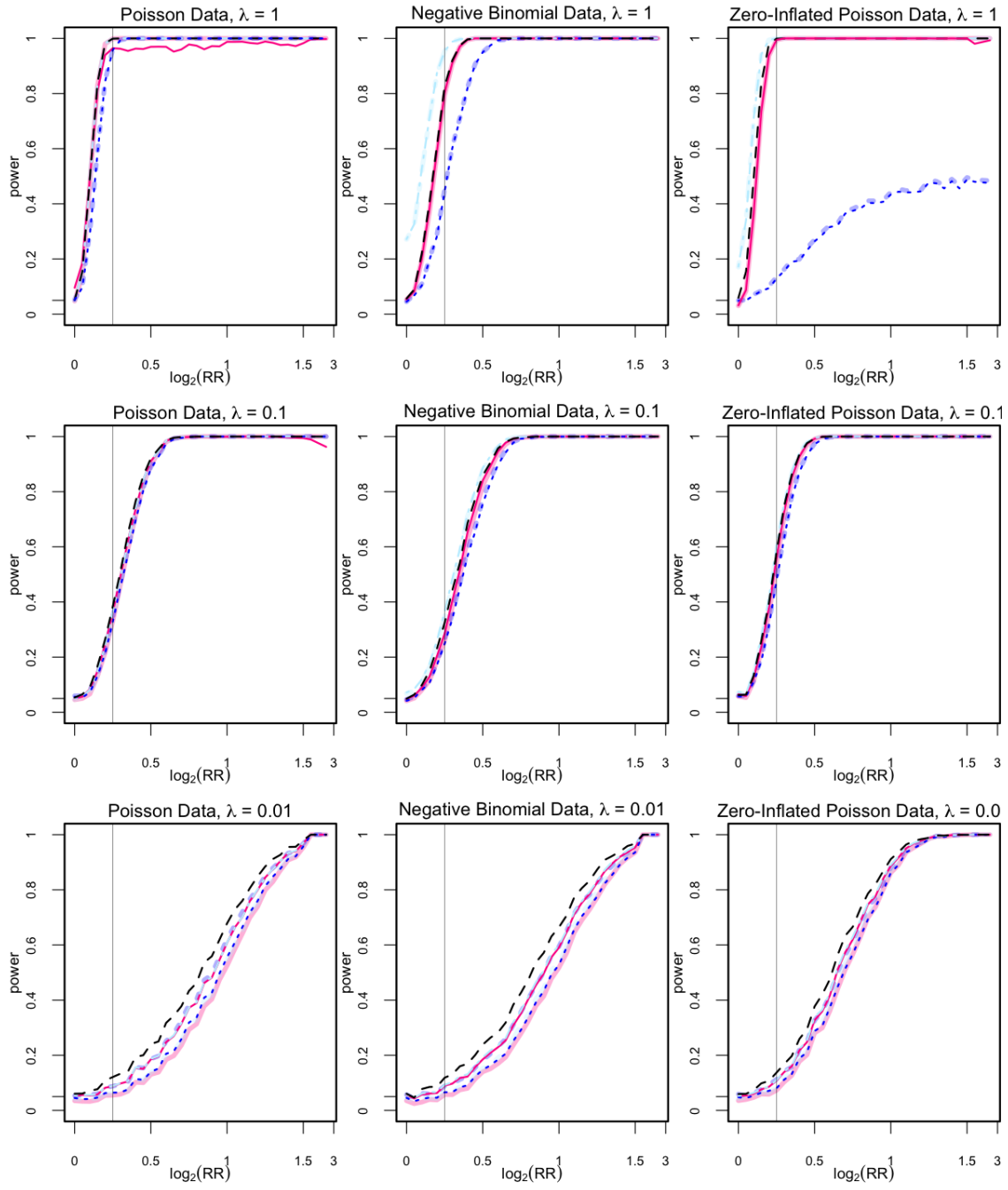

Figure A.7: Power of all CPT code-specific tests with **unequal** sample sizes ($n_0 = 1000$, $n_1 = 3000$) using Negative Binomial data, Poisson data, or Zero-inflated Poisson data, each with a mean of $\lambda_0 = 1$ or 0.01 in cohort 0, and a rate ratio on log 2 scale ($\log_2 \frac{\lambda_1}{\lambda_0}$) ranging from 0 to 1.5. Colored lines correspond to Negative Binomial LRT (——); Negative Binomial ET (——); Poisson LRT (- - -); Poisson ET (- - -); Fisher's ET (- - -); Binomial LRT (- - -); t-test (— —).

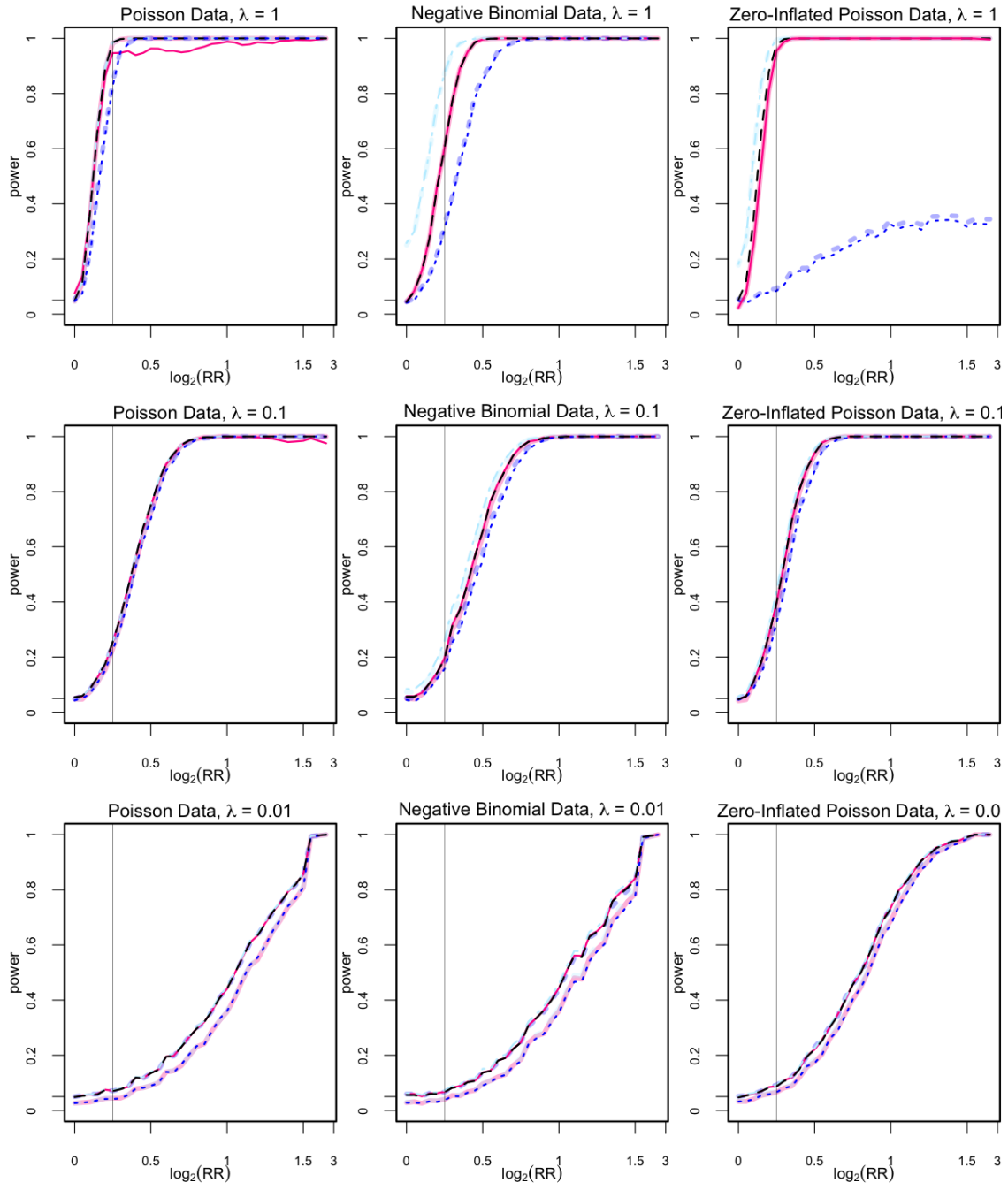

Figure A.8: Power of all CPT code-specific tests with **equal** sample sizes ($n_0 = 1000$, $n_1 = 1000$) using Negative Binomial data, Poisson data, or Zero-inflated Poisson data, each with a mean of $\lambda_0 = 1$ or 0.01 in cohort 0, and a rate ratio on log 2 scale ($\log_2 \frac{\lambda_1}{\lambda_0}$) ranging from 0 to 1.5. Colored lines correspond to Negative Binomial LRT (——); Negative Binomial ET (——); Poisson LRT (- - -); Poisson ET (- - -); Fisher's ET (- - -); Binomial LRT (- - -); t-test (— —).

Appendix B

APPENDIX FOR CHAPTER 3

B.1 Proof of Lemmas and Theorems in Section 3.3.5

Lemma 1 [Adjustment Term (B)]

$$E\left[\frac{\partial}{\partial\theta}\big\{g_1(h_*(X),\theta)-g_0(h_*(X),\theta)\big\}\right]=E[\big(\frac{T}{h_*(X)}(Y-g_1(h_*(X)))-\frac{1-T}{1-h_*(X)}(Y-g_0(h_*(X)))\big)\cdot S(O)]$$

The adjustment term is

$$\frac{T}{h_*(X)}(Y-g_1(h_*(X)))-\frac{1-T}{1-h_*(X)}(Y-g_0(h_*(X))) \tag{B.1}$$

Proof. Since

$$\begin{cases} g_1(s_*,\theta) & =E_\theta[Y|T=1,S_*=s_*] \\ g_0(s_*,\theta) & =E_\theta[Y|T=0,S_*=s_*], \end{cases}$$

let $t_1(S_*)$ and $t_2(S_*)$ denote any function of S_*, we have the following orthogonally condition:

$$E_\theta[T\cdot\big(Y-g_1(S_*,\theta)\big)\cdot t_1(S_*)]=E_\theta[T\cdot\big(E_\theta[Y|T=1,S_*]-g_1(S_*,\theta)\big)\cdot t_1(S_*)]=0$$

Analogously,

$$E_\theta[(1-T)\cdot\big(Y-g_0(S_*,\theta)\big)\cdot t_2(S_*)]=E_\theta[(1-T)\cdot\big(E_\theta[Y|T=0,S_*]-g_0(S_*,\theta)\big)\cdot t_2(S_*)]=0$$

Thus,

$$\begin{cases} E_\theta[T\cdot\big(Y-g_1(S_*,\theta)\big)\cdot t_1(S_*)]=0 \\ E_\theta[(1-T)\cdot\big(Y-g_0(S_*,\theta)\big)\cdot t_2(S_*)]=0. \end{cases}$$

Let $t_1(S_*) = \frac{1}{S_*}$ and $t_2(S_*) = \frac{1}{1-S_*}$, then we have

$$\begin{cases} E_\theta[\frac{TY}{S_*}] = E_\theta[\frac{Tg_1(S_*,\theta)}{S_*}] \\ E_\theta[\frac{(1-T)Y}{1-S_*}] = E_\theta[\frac{(1-T)g_0(S_*,\theta)}{1-S_*}]. \end{cases}$$

Taking derivative on both sides, we have

$$\begin{aligned} \text{Left} &= \frac{\partial}{\partial\theta} E_\theta[\frac{TY}{S_*}] \\ &= E[\frac{TY}{S_*} \cdot S(O)] \end{aligned}$$

and

$$\begin{aligned} \text{Right} &= \frac{\partial}{\partial\theta} E_\theta[\frac{Tg_1(S_*,\theta)}{S_*}] \\ &= \frac{\partial}{\partial\theta} E_\theta[\frac{Tg_1(S_*)}{S_*}] + \frac{\partial}{\partial\theta} E[\frac{Tg_1(S_*,\theta)}{S_*}] \\ &= E_\theta[\frac{Tg_1(S_*)}{S_*} \cdot S(O)] + \frac{\partial}{\partial\theta} E[\frac{E[T|X]g_1(S_*,\theta)}{S_*}] \\ &= E_\theta[\frac{Tg_1(S_*)}{S_*} \cdot S(O)] + \frac{\partial}{\partial\theta} E[g_1(S_*,\theta)] \end{aligned}$$

Combining the above, we have

$$\frac{\partial}{\partial\theta} E[g_1(S_*,\theta)] = E_\theta[\frac{T\big(Y - g_1(S_*)\big)}{S_*} \cdot S(O)] \tag{B.2}$$

Analogously,

$$\frac{\partial}{\partial\theta} E[g_0(S_*,\theta)] = E_\theta[\frac{(1-T)\big(Y - g_0(S_*)\big)}{1-S_*} \cdot S(O)] \tag{B.3}$$

Combine (B.2) and (B.3) we get

$$\frac{\partial}{\partial\theta} E\left[\big\{g_1(h_*(X),\theta) - g_0(h_*(X),\theta)\big\}\right] = E[\big(\frac{T}{h_*(X)}(Y - g_1(h_*(X))) - \frac{1-T}{1-h_*(X)}(Y - g_0(h_*(X)))\big) \cdot S(O)]$$

□

Lemma 2 [Adjustment Term (C)]

$$\frac{\partial}{\partial\theta} E\left[\big\{g_1(h_*(X,\theta)) - g_0(h_*(X,\theta))\big\}\right] = E[\delta(X) \cdot \big(T - h_*(X)\big) \cdot S(O)], \tag{B.4}$$

where

$$\delta(X) = (-1)\cdot\big(\frac{E[Y|T=1,X]-E[Y|T=1,h_*(X)]}{h_*(X)}+\frac{E[Y|T=0,X]-E[Y|T=0,h_*(X)]}{1-h_*(X)}\big). \tag{B.5}$$

Proof. Recall that $S_\theta = h_*(X,\theta) = E_\theta[T|X]$, with $S_\theta|_{\theta=0} = S_*$. So $\frac{\partial}{\partial\theta}E\left[\{g_1(h_*(X,\theta)) - g_0(h_*(X,\theta))\}\right] = \frac{\partial}{\partial\theta}E\left[\{g_1(S_\theta) - g_0(S_\theta)\}\right]$.

Since

$$\begin{cases} g_1(S_\theta) & = E[Y|T=1,S_\theta] \\ g_0(S_\theta) & = E[Y|T=0,S_\theta], \end{cases}$$

let $t_1(S_\theta)$ and $t_2(S_\theta)$ denote any function of S_θ, we have the following orthogonally condition:

$$E[T\cdot\big(Y-g_1(S_\theta)\big)\cdot t_1(S_\theta)] = E[T\cdot\big(E_\theta[Y|T=1,S_\theta]-g_1(S_\theta)\big)\cdot t_1(S_\theta)] = 0$$

Analogously,

$$E[(1-T)\cdot\big(Y-g_0(S_\theta)\big)\cdot t_2(S_\theta)] = E[(1-T)\cdot\big(E_\theta[Y|T=0,S_\theta]-g_0(S_\theta)\big)\cdot t_2(S_\theta)] = 0$$

Thus,

$$\begin{cases} E[T\cdot\big(Y-g_1(S_\theta)\big)\cdot t_1(S_\theta)] = 0 \\ E[(1-T)\cdot\big(Y-g_0(S_\theta)\big)\cdot t_2(S_\theta)] = 0. \end{cases}$$

Let $t_1(S_\theta) = \frac{1}{S_\theta}$ and $t_2(S_\theta) = \frac{1}{1-S_\theta}$, then we have

$$\begin{cases} E[\frac{TY}{S_\theta}] = E[\frac{Tg_1(S_\theta)}{S_\theta}] \\ E[\frac{(1-T)Y}{1-S_\theta}] = E[\frac{(1-T)g_0(S_\theta)}{1-S_\theta}]. \end{cases}$$

which are equivalent to

$$\begin{cases} E[\frac{E[TY|X]}{S_\theta}] = E[\frac{S_*E[Y|T=1,X]}{S_\theta}] = E[\frac{E[T|X]g_1(S_\theta)}{S_\theta}] = E[\frac{S_*g_1(S_\theta)}{S_\theta}] \\ E[\frac{E[(1-T)Y|X]}{1-S_\theta}] = E[\frac{(1-S_*)E[Y|T=0,X]}{S_\theta}] = E[\frac{E[1-T|X]g_0(S_\theta)}{1-S_\theta}] = E[\frac{(1-S_*)g_0(S_\theta)}{1-S_\theta}]. \end{cases}$$

Taking derivative on both sides, we have

$$\begin{aligned}\text{Left} &= \frac{\partial}{\partial\theta}E[\frac{S_* E[Y|T=1,X]}{S_\theta}] \\ &= E[\frac{-1}{S_\theta^2}|_{\theta=0}\cdot S_*\cdot E[Y|T=1,X]\cdot\frac{\partial S_\theta}{\partial\theta}] \\ &= E[\frac{-1}{S_*}\cdot E[Y|T=1,X]\cdot\frac{\partial S_\theta}{\partial\theta}]\end{aligned}$$

and

$$\begin{aligned}\text{Right} &= \frac{\partial}{\partial\theta}E[\frac{S_* g_1(S_\theta)}{S_\theta}] \\ &= E[\frac{-1}{S_\theta^2}|_{\theta=0}\cdot S_*\cdot g_1(S_\theta)\cdot\frac{\partial S_\theta}{\partial\theta}] + E[\frac{S_*}{S_\theta}|_{\theta=0}\cdot\frac{\partial}{\partial\theta}g_1(S_\theta)] \\ &= E[\frac{-1}{S_*}\cdot g_1(S_*)\cdot\frac{\partial S_\theta}{\partial\theta}] + E[\frac{\partial}{\partial\theta}g_1(S_\theta)]\end{aligned}$$

Combining the above, we have

$$E[\frac{\partial}{\partial\theta}g_1(S_\theta)] = E[\frac{-1}{S_*}\cdot\big(E[Y|T=1,X]-E[Y|T=1,h_*(X)]\big)\cdot\frac{\partial}{\partial\theta}S_\theta] \tag{B.6}$$

Analogously,

$$E[\frac{\partial}{\partial\theta}g_0(S_\theta)] = E[\frac{-1}{1-S_*}\cdot\big(E[Y|T=0,X]-E[Y|T=0,h_*(X)]\big)\cdot\frac{\partial}{\partial\theta}S_\theta] \tag{B.7}$$

Combine (B.6) and (B.7) we get

$$\begin{aligned}&\frac{\partial}{\partial\theta}E\left[\big\{g_1(h_*(X,\theta))-g_0(h_*(X,\theta))\big\}\right] \\ &= \frac{\partial}{\partial\theta}E[(-1)\big(\frac{E[Y|T=1,X]-E[Y|T=1,h_*(X)]}{S_*}-\frac{E[Y|T=0,X]-E[Y|T=0,h_*(X)]}{1-S_*}\big)S_\theta]\end{aligned}$$

Let $\delta(X) = (-1) \cdot \big(\frac{E[Y|T=1,X]-E[Y|T=1,h_*(X)]}{h_*(X)} - \frac{E[Y|T=0,X]-E[Y|T=0,h_*(X)]}{1-h_*(X)}\big)$, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta} E\left[\left\{g_1(h_*(X,\theta)) - g_0(h_*(X,\theta))\right\}\right] \\
&= \frac{\partial}{\partial\theta} E[\delta(X) \cdot E_\theta[T|X]] \\
&= \frac{\partial}{\partial\theta} E_\theta[\delta(X) \cdot E_\theta[T|X]] - \frac{\partial}{\partial\theta} E_\theta[\delta(X) \cdot E[T|X]] \\
&= \frac{\partial}{\partial\theta} E_\theta[\delta(X) \cdot d] - \frac{\partial}{\partial\theta} E_\theta[\delta(X) \cdot E[T|X]] \\
&= \frac{\partial}{\partial\theta} E_\theta[\delta(X) \cdot (T - E[T|X])] \\
&= E[\delta(X) \cdot (T - E[T|X]) \cdot S(O)]
\end{aligned}$$

□

Theorem 1:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\beta(\theta) &= E[\big(g_1(h_*(X)) - g_0(h_*(X)) - \beta_*\big)S(O)] \\
&+ E[\big(\frac{T}{h_*(X)}(Y - g_1(h_*(X))) - \frac{1-T}{1-h_*(X)}(Y - g_0(h_*(X)))\big) \cdot S(O)] \\
&+ E[\delta(X) \cdot \big(T - h_*(X)\big) \cdot S(O)] \\
&= E[\big\{ \\
&\quad E[Y|T=1,X] - E[Y|T=0,X] - \beta_* \\
&+ \frac{T}{h_*(X)}(Y - E[Y|T=1,X]) - \frac{1-T}{1-h_*(X)}(Y - E[Y|T=0,X]) \\
&\quad \big\} \cdot S(O)]
\end{aligned}$$

Thus, the influence function is

$$E[Y|T=1,X] - E[Y|T=0,X] - \beta_* + \frac{T}{h_*(X)}(Y - E[Y|T=1,X]) - \frac{1-T}{1-h_*(X)}(Y - E[Y|T=0,X]), \tag{B.8}$$

which is equal to the EIF Equation (3.8).

Proof. The influence function is simply the sum of (3.26), (B.1), and (3.28), which is

$$E[Y|T=1,X]-E[Y|T=0,X]-\beta_*+\frac{T}{h_*(X)}(Y-E[Y|T=1,X])-\frac{1-T}{1-h_*(X)}(Y-E[Y|T=0,X]). \tag{B.9}$$

□

B.2 Regularity Conditions in Mammen (2016)

Assumption 1: The sample observations $O_i = (Y_i, T_i, X_i), i = 1, \ldots, n$ are i.i.d.

Assumption 2:

(i) The random vetor X is continuously distributed with compact support I_X. Its density function f_X is bounded and bounded away from zero on I_X, and is also $q+1$ times continuously differentiable for some uneven number $q >$dim(X).

(ii) The function $h(X)$ is bounded away from zero and one on I_X, and is also $q+1$ times continuously differentiable.

(iii) For any $t \in \{0,1\}$, the random variable $S = P(X)$ is continuously distributed conditional on $T = t$, with compact support I_S. Its conditional density function $f_{P|T}(\cdot, t)$ is bounded and bounded away from zero on I_S, and is also four times continuously differentiable.

(iv) For any $t \in \{0,1\}$, the function $g(s)$ is four times continuously differentiable on I_S.

Assumption 3: The residual $\epsilon = Y - E[Y|S = P(X)]$ satisfies $E[\exp l \cdot |\epsilon| \mid X] \leq C$ almost surely for a constant $C > 0$ and $l > 0$ small enough.

Assumption 4:

(i) The kernel K is twice continuously differentiable and satisfies the following conditions: $\int K(u)du = 1, \int uK(u)du = 0, \int |u^2K(u)|du < \infty$, and $K(u) = 0$ for values of u not contained in some compact interval, say $[-1, 1]$.

(ii) The kernel $\mathcal{L}$ is k-times continuously differentiable for some natural number $k \geq \max\{2, \dim(X)/2\}$, and satisfies the following conditions:$\int \mathcal{L}(u)du = 1, \int u\mathcal{L}(u)du = 0$ and $\mathcal{L}(u) = 0$ for values of u not contained in some compact interval, say $[-1, 1]$.

Assumption 5: The bandwidth $h_S \sim n^{-\gamma}$ for $q-$th order local polynomial smoothing in estimation of $S = P(X)$ satisfies $\gamma = 1/(2q + 1)$; the bandwidth $h_g \sim n^{-\eta}$ for third order local polynomial smoothing in estimation of $E[Y|T = t, S] = g_t(S)$ satisfies $1/8 < \eta < (q + 2)/(8q + 4)$.

B.3 Regularity Conditions and Proof Extending Hahn (2016)

In this section, we first list key notation, provide definitions of Riesz representers and a list of regularity conditions (Assumptions 1.1 - 1.4), then provide proofs of Theorem 3.2 and Corollary 3.1.

Key Notation

Observations: $O = (X, Y, T)$

A vector of basis functions: $R(x)_{L\times 1} = [r_1(x), \ldots, r_L(x)]^T$; $B[h]_{k\times 1} = [p_1(h), \ldots, p_K(h)]^T$, where $L = L(n)$ and $K = K(n)$ are number of basis functions that grows with sample size n. The matrix $R_n = [R(X_1), \ldots, R(X_n)]$ with $\dim(R_n)=L \times n$, and $\hat{B}_n = [B(\hat{h}_1), \ldots, B(\hat{h}_n)]^T$, $\hat{h}_i = \hat{h}_n(X_i)$ with $\dim(\hat{B}_n)=n \times K$. $Q_L = E[R(X)R(X)^T]$ and $Q_K = E[B(h(X))B(h(X))^T]$. $Q_{n,L} = \frac{1}{n}R_nR_n^T$ and $\hat{Q}_{n,K} = \frac{1}{n}\hat{B}_n^T\hat{B}_n = \frac{1}{n}B(\hat{h}_n)^TB(\hat{h}_n)$.

Supports:

(1) Support of h: $h_\eta = [-\eta, 1+\eta]$ for some small $\eta > 0$

(2) Support of X: $\mathcal{X}$.

Bounds:

(1) $\nu_{j,k} = \sup_{h \in h_\eta} \|\partial^j B(h)^T \beta_{0,K}\|, j = 1, 2$, where $\beta_{0,K}$ is defined in Assumption 1.2 (iii)

(2) $\zeta_{0,L}$ such that $\sup_{x \in \mathcal{X}} \left| R(X) \right| \leq \zeta_{0,L}$

(3) $\xi_{j,K}$ such that $\sup_{h \in h_\eta} \left| \partial^j B(h) \right| \leq \xi_{j,K}$

(4) Generic constant C.

Norms:

(1) ℓ-2 norm: $\|\cdot\|$ or $\|\cdot\|_2$. For a matrix A, $\|A\|_2 = \max \sqrt{\lambda(A^T A)}$ is also called the spectral norm.

(2) $\omega_{\max}(A), \omega_{\min}(A)$: largest or smallest eigenvalue of a symmetric matrix A.

(3) For a function $f(x)$, $|f|_d = \max_{|\lambda| \leq d} \sup_{x \in \mathcal{X}} \left| \partial^\lambda f(x) \right|$. For example, $|g|_d = \max_{|\lambda| \leq d} \sup_{h \in h_\eta} \left| \partial^\lambda g(h) \right|$.

(4) Uniform norm: $\|\cdot\|_\infty = |f|_{d=0} = \sup_{x \in \mathcal{X}} \left| f(x) \right|$. For example, $\|g\|_\infty = |g|_{d=0} = \sup_{h \in h_\eta} \left| g(h) \right|$.

Functional derivatives:

(1) $\partial B[\tilde{h}(X_i)] = \frac{\partial B(h(X_i))}{\partial h}\big|_{h=\tilde{h}}$

(2) $\partial g_0(h_0) = \frac{\partial g_0(h)}{\partial h}\Big|_{h=h_0}$ and $\partial^2 g_0(\tilde{h}) = \frac{\partial^2 g_0(h)}{\partial h^2}\Big|_{h=\tilde{h}}$

(3) $\partial \rho_g(g) = \frac{\partial \rho(h_0, g+\tau v)}{\partial \tau}\Big|_{\tau=0}$ and $\partial \rho_g(g_0) = \frac{\partial \rho(h_0, g_0+\tau v)}{\partial \tau}\Big|_{\tau=0}$ for all $v \in \mathcal{V}_2$

(4) $\triangle_\varphi(O, h) = T - h(X)$ and $\triangle_\psi(O, g, h) = \frac{T}{h}[Y - g(h(X))]$

Riesz Representers

We use $\triangle_\varphi(O, h_0)[h - h_0]$ to denote the linear approximation of $\varphi(O, h) - \varphi(O, h_0)$ (up to a second order term) such that $\triangle_\varphi(O, h_0)[h - h_0]$ is linear in $h - h_0$. In fact, since $\varphi(O, h) = \frac{-1}{2}[T - h(X)]^2$, we have $\triangle_\varphi(O, h) = T - h(X)$. Because h_0 is unique, we define a norm $\|\cdot\|_\varphi$ on $\mathcal{N}_{h,n}$ as

$$\|h - h_0\|_\varphi^2 \equiv -\frac{\partial}{\partial\tau}\Big|_{\tau=0} E\left\{\triangle_\varphi(O, h_0 + \tau[h - h_0])[h - h_0]\right\} = E\{[h - h_0]^2\}.$$

Let $\mathcal{V}_1$ be the closed linear span of $\mathcal{N}_{h,n} - \{h_0\}$ under $\|\cdot\|_\varphi$ which is a Hilbert space under $\|\cdot\|_\varphi$ with the corresponding inner product $<\cdot,\cdot>_\varphi$ defined as

$$< v_{h_1}, v_{h_2} >_\varphi \equiv -\frac{\partial}{\partial\tau}\Big|_{\tau=0} E\left\{\triangle_\varphi(O, h_0 + \tau v_{h_2})[v_{h_1}]\right\} = E[v_{h_1} v_{h_2}]$$

for any $v_{h_1}, v_{h_2} \in \mathcal{V}_1$.

We assume that there is a linear functional $\partial_h\rho(g_0, h_0)[\cdot] : \mathcal{V}_1 \to R$ such that

$$\partial_h\rho(g_0, h_0)[v] = \frac{\partial}{\partial\tau}\Big|_{\tau=0}\rho(h_0 + \tau v, g_0) \text{ for all } v \in \mathcal{V}_1.$$

Let $h_{0,n}$ denote the project of h_0 on $\mathcal{H}_n$ under the norm $\|\cdot\|_\varphi$. Let $\mathcal{V}_{1,n}$ denote the Hilbert space generated by $\mathcal{N}_{h,n} - \{h_{0,n}\}$. Then $\dim(\mathcal{V}_{1,n}) = L(n) < \infty$. By Riesz representation theorem, there is a sieve Riesz representer $d_{h_n}^* \in \mathcal{V}_{1,n}$ such that

$$\partial_h\rho(g_0, h_0)[v] = < d_{h_n}^*, v >_\varphi \text{ for all } v \in \mathcal{V}_{1,n}, \text{ and } \|d_{h_n}^*\|_\varphi^2 = \sup_{0 \neq v \in \mathcal{V}_{1,n}} \frac{|\partial_h\rho(g_0, h_0)[v]|^2}{\|v\|_\varphi^2}.$$

Moreover, the linear functional $\partial_h\rho(g_0, h_0)[\cdot] : \mathcal{V}_1 \to R$ is bounded if and only if $\lim_{L(n)\to\infty} \|d_{h_n}^*\|_\varphi < \infty$.

Similarly, we use $\triangle_\psi(O, g_0, h_0)[g - g_0]$ to denote the linear approximation of $\varphi(O, g, h_0) - \varphi(O, g_0, h_0)$ (up to a second order term) such that $\triangle_\psi(O, g_0, h_0)[g - g_0]$ is linear in $g - g_0$. In fact, since $\psi(O, g, h) = \frac{-1}{2}\frac{T}{h}[Y - g(h)]^2$, we have $\triangle_\psi(O, g, h) = \frac{T}{h}[Y - g(h(X))]$. Because g_0 is unique, we define a norm $\|\cdot\|_\psi$ on $\mathcal{N}_{g,n}$ as

$$\|g - g_0\|_\varphi^2 \equiv -\frac{\partial}{\partial\tau}\Big|_{\tau=0} E\left\{\triangle_\psi(O, g_0 + \tau[g - g_0], h_0)[g - g_0]\right\} = E\{1(T = 1)[g(h_0) - g_0(h_0)]^2\}.$$

Let $\mathcal{V}_2$ be the closed linear span of $\mathcal{N}_{g,n} - \{g_0\}$ under $\|\cdot\|_\psi$ which is a Hilbert space under $\|\cdot\|_\psi$ with the corresponding inner product $<\cdot,\cdot>_\psi$ defined as

$$< v_{g_1}, v_{g_2} >_\psi \equiv -\frac{\partial}{\partial\tau}\Big|_{\tau=0} E\left\{\triangle_\psi(O, g_0 + \tau v_{g_2}, h_0)[v_{g_1}]\right\} = E[1(T=1)v_{g_1}v_{g_2}]$$

for any $v_{g_1}, v_{g_2} \in \mathcal{V}_2$.

We assume that there is a linear functional $\partial_g\rho(g_0, h_0)[\cdot] : \mathcal{V}_2 \to R$ such that

$$\partial_g\rho(g_0, h_0)[v] = \frac{\partial}{\partial\tau}\Big|_{\tau=0}\rho(h_0, g_0 + \tau v) \text{ for all } v \in \mathcal{V}_2. \tag{B.10}$$

Let $g_{0,n}$ denote the project of g_0 on $\mathcal{G}_n$ under the norm $\|\cdot\|_\psi$. Let $\mathcal{V}_{2,n}$ denote the Hilbert space generated by $\mathcal{N}_{g,n} - \{g_{0,n}\}$. Then $\dim(\mathcal{V}_{2,n}) = K(n) < \infty$. By Riesz representation theorem, there is a sieve Riesz representer $d^*_{g_n} = d^*_{g_n}(h_0) \in \mathcal{V}_{2,n}$ such that

$$\partial_g\rho(g_0, h_0)[v] =< d^*_{g_n}, v >_\psi \text{ for all } v \in \mathcal{V}_{2,n}, \text{ and } \|d^*_{g_n}\|^2_\varphi = \sup_{0\neq v\in\mathcal{V}_{2,n}} \frac{|\partial_g\rho(g_0, h_0)[v]|^2}{\|v\|^2_\psi}.$$

Moreover, the linear functional $\partial_g\rho(g_0, h_0)[\cdot] : \mathcal{V}_2 \to R$ is bounded if and only if $\lim_{K(n)\to\infty} \|d^*_{g_n}\|_\psi < \infty$.

Because $\mathcal{V}_{2,n}$ is the linear space spanned by the basis functions $B(h_0)$, the Riesz representer $d^*_{g_n}$ of the functional $\partial\rho_g(g_0)[\cdot]$ has a closed form expression

$$d^*_{g_n}(h_0) = \partial\rho_g(g_0)[B]^T Q^{-1}_{K(n)} B(h_0),$$

where $\partial\rho_g(g_0)[B]^T = [\partial\rho_g(g_0)[p_1], \ldots, \partial\rho_g(g_0)[p_{K(n)}]]$ and $Q_{K(n)} = E[B(h_0)B(h_0)^T]$. By Eq. (B.10) we have

$$d^*_{g_n}(h_0) = E[B(h_0)]^T Q^{-1}_{K(n)} B(h_0).$$

In addition, we define

$$\Gamma_g(h_0, g_0)[v_g] = \frac{\partial}{\partial\tau}\Big|_{\tau=0} E[\psi(O, g_0 + \tau v_g, h_0)] \text{ for any } v_g \in \mathcal{V}_2$$

and

$$\Gamma(h_0, g_0)[v_h, v_g] = \frac{\partial}{\partial\tau}\Big|_{\tau=0} \Gamma_g(h_0 + \tau v_h, g_0)[v_g] \text{ for any } v_h \in \mathcal{V}_1,$$

then $\Gamma(h_0, g_0)[v_h, v_g]$ is a bilinear functional on $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$ which describes the effect of the first-step estimation on the asymptotic behavior of the second-step estimator. We define $d^*_{\Gamma_n} \in \mathcal{V}_{1,n}$ as

$$\Gamma(h_0, g_0)[v_h, d^*_{g_n}] =< v_h, d^*_{\Gamma_n} >_\varphi \text{ for any } v_h \in \mathcal{V}_{1,n}.$$

Because $\mathcal{V}_{1,n}$ is the linear space spanned by the basis functions $R(x)$, the Riesz representer $d^*_{\Gamma_n}$ of the functional $\Gamma(h_0, g_0)[v_h, d^*_{g_n}]$ has a closed form expression

$$d^*_{\Gamma_n}(x) = E[\partial g_0(h_0) d^*_{g_n}(h_0) R(X)^T] Q^{-1}_{L(n)} R(x),$$

where $Q_{L(n)} = E[R(X)R(X)^T]$.

Using the sieve Riesz representers $d^*_{h_n}$, $d^*_{g_n}$, and $d^*_{\Gamma_n}$ we define

$$\|v^*_n\|^2_{sd} = \text{Var}[\frac{1}{n}\sum_{i=1}^{n}\Big(g_0(h_0(X_i)) - \rho(g_0(h_0)) + \triangle_\varphi(O_i, h_0)[d^*_{h_n}] + \triangle_\varphi(O_i, h_0)[d^*_{\Gamma_n}] + \triangle_\psi(O_i, g_0, h_0)[d^*_{g_n}]\Big)].$$

In addition, we define normalized Riesz representer $(u^*_{g_n}, u^*_{h_n}, u^*_{\Gamma_n}) = \frac{1}{\|v^*_n\|_{sd}}(d^*_{g_n}, d^*_{h_n}, d^*_{\Gamma_n})$

Assumptions

Assumption 1.1

(i) $\{Y_i, T_i, X_i\}_{i=1}^n$ are independent and identically distributed

(ii) $E[\epsilon_i^4 | X_i] < C$ and $E[\epsilon_i^2 | X_i] > C^{-1}$

(iii) There exists $\rho_h > 0$ and $\gamma_{0,L} \in R^L$ such that for an integer $d > 0$,

$$|h_0 - h_{0,L}|_d = O(L^{-\rho_h}) \quad \text{as } L \to \infty,$$

where $h_{0,L} = R(\cdot)^T \gamma_{0,L}$.

(iv) For all L, the eigenvalues of $Q_L = E[R(X)R(X)^T]$ are between C^{-1} and C.

(v) There exists a nondecreasing sequence $\zeta_{0,L}$ such that

$$\|R(X)\|_\infty = \zeta_{0,L}$$

Comments:

1. For Assumption 1.1 (iii), by [72] and [31], for regression splines,

$$\rho_h = s/r \quad \text{when } d = 0$$
$$\rho_h = s - d \quad \text{when } r = 1 \quad \text{and } d \neq 0$$

where s is the number of continuous derivatives of $h_0(x)$ that exist, r is the dimension of X. In our case,

2. By [72], for any integer $d \geq 0$, let $\xi_{d,L}$ such that $\max_{|\lambda| \leq d} \sup_{x \in \mathcal{X}} \left| \partial^j R(x) \right| = |R(x)|_d \leq \xi_{d,L}$, then for B-splines, we have $\xi_{d,L} = L^{\frac{1}{2}+d}$. Therefore $\zeta_{0,L} = L^{\frac{1}{2}}$.

Assumption 1.2

(i) $E[u_i^4|X_i] < C$ and $E[u_i^2|X_i] > C^{-1}$

(ii) There exists $\rho_g > 0$ and $\beta_{0,K} \in R^K$ such that for an integer $d > 0$,

$$|g_0 - g_{0,K}|_d = O(K^{-\rho_g}) \quad \text{as } K \to \infty,$$

where $g_{0,K} = B(\cdot)^T \beta_{0,K}$.

(iii) $g_0(h)$ is twice continuously differentiable.

(iv) For all K, the eigenvalues of $Q_K = E[B(h(X))B(h(X))^T]$ are between C^{-1} and C.

(v) For each of $d = 0, 1, 2$, there exists a nondecreasing (w.r.t K) sequence $\xi_{d,K}$ such that

$$\sup_{h \in h_\eta} \|\partial^d B(h)\| \leq \xi_{d,K},$$

Comments:

Similar to Assumption 1.1, by [72], for B-splines, we have $\xi_{d,K} = K^{\frac{1}{2}+d}$ for $d = 0, 1, 2$.

Assumption 1.3

(i) The Riesz Representer $d^*_{g_n}$ satisfies $\|d^*_{g_n}\| \geq C$ for all n.

(ii) $\rho(g) = \int_{\mathcal{X}} g dP(x)$ satisfies

$$\sup_{g \in \mathcal{N}_{g,n}} \left| \frac{\rho(g) - \rho(g_0) - \partial(g_0)[g - g_0]}{\|v^*_n\|_{sd}} \right| = o(n^{-1/2})$$

(iii) $\left| \|v^*_n\|^{-1}_{sd} \partial\rho_g(g_0)[g_{0,n} - g_0] \right| = o(n^{-1/2})$

(iv) $\sup_{g \in \mathcal{N}_{g,n}} \|\partial\rho_g(g)[P] - \partial\rho_g(g_0)[P]\| = o(1)$

Comments:

Easy to verify for linear functional $\rho(g) = \int_{\mathcal{X}} g dP(x)$.

Assumption 1.4

(i) $n^{-1/2}(K+L)^{\frac{1}{2}}(\xi_{0,K} + \zeta_{0,L})(\log n)^{\frac{1}{2}} = o(1)$

(ii) $n^{-1}(L\xi^2_{1,K} \log n + \zeta_{0,L}\xi_{1,K}) = o(1)$

(iii) $n^{-1/2}(L\xi_{2,K} + L^{\frac{1}{2}}\xi_{1,K})(n^{-1/2}K^{\frac{1}{2}} + K^{-\rho_g} + \nu_{1,K}n^{-1/2}L^{\frac{1}{2}}) \log n = o(1)$

(iv) $n^{-1/2}(L + L^{\frac{1}{2}}\nu_{1,K} + L\nu_{2,K}) \log n = o(1)$

(v) $nL^{1-2\rho_h} + K^{-\rho_g} = o(1)$

Comments:

Assumption 1.4 implies that

1. For propensity score model, $L \propto n^{\frac{1}{2\rho_h - 1}}$
2. For outcome regression model, $\frac{(\log n)^2}{n} L^3 K^5 = o(1)$

In practice we recommend

1. L: undersmoothing compared to traditional nonparametric rates $n^{\frac{1}{2\rho_h + 1}}$

2. K: small, e.g. K=3 (cubic spline) for sufficiently large ρ_s and sample size n

Lemma B.1 (Rate of Convergence)**.** *Under Assumptions 1.1 and 1.4(i), we have*

$$\|\hat{h}_n - h_0\|_2 = O_p(\delta^*_{h,n}),$$

where $\delta^*_{h,n} = L^{1/2}n^{-1/2} + L^{-\rho_h}$.

Proof. Recall that $\hat{h}_n(\cdot) = R(\cdot)^T\hat{\gamma}_n$. By the triangle inequality, we obtain

$$\begin{aligned}
&\|\hat{h}_n - h_0\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}|\hat{h}_n(X_i) - h_{0,L}(X_i)|^2 + \frac{1}{n}\sum_{i=1}^{n}|h_{0,L}(X_i) - h_0(X_i)|^2 \\
&\leq (\underset{1\times L}{\hat{\gamma}_n - \gamma_{0,L}})^T \underset{L\times L}{Q_{n,L}} (\underset{L\times 1}{\hat{\gamma}_n - \gamma_{0,L}}) + O(L^{-2\rho_h}) \quad \text{by Assumption 1.1(iii)} \\
&\leq \omega_{\max}(Q_{n,L})\|\hat{\gamma}_n - \gamma_{0,L}\|^2 + O(L^{-2\rho_h}),
\end{aligned} \tag{B.11}$$

where the last inequality holds by the Holder's inequality. It suffices to obtain upper bounds for $\omega_{\max}(Q_{n,L})$ and $\|\hat{\gamma}_n - \gamma_{0,L}\|$.

To obtain an upper bound for $\omega_{\max}(Q_{n,L})$, we need to control the difference between $Q_{n,L}$ and Q_L under the spectral norm. By Lemma 6.2 of [11], we have

$$\|Q_L - Q_{n,L}\| = O_p(\zeta_{0,L}\sqrt{\log(n)/n}) \tag{B.12}$$

which is $o(1)$ by Assumption 1.4(i). By Assumption 1.1(iv) we have

$$(2C)^{-1} < \omega_{\min}(Q_{n,L}) < \omega_{\max}(Q_{n,L}) < 2C \ \text{ wpa1.} \tag{B.13}$$

By Assumptions 1.1 and 1.4(i), along with Theorem 4.1 in [11], we have

$$\|\hat{\gamma}_n - \gamma_{0,L}\| = O_p(\delta^*_{h,n}). \tag{B.14}$$

Finally, combining (B.11), (B.13), and (B.14), we have

$$\|\hat{h}_n - h_0\| = O_p(\delta^*_{h,n}).$$

□

Lemma B.2 (Rate of Convergence)**.** *Under Assumptions 1.2 (ii) and (iv), 1.4(i), (ii) and (v), we have*

$$\|\hat{g}_n(\hat{h}_n) - g_0(h_0)\|_2 = O_p(\delta^*_{g,n}),$$

where $\delta^*_{g,n} = K^{1/2}n^{-1/2} + K^{-\rho_g} + \nu_{1,K}\delta^*_{h,n}$.

Proof. By the triangle inequality, we obtain

$$\begin{aligned}
&\|\hat{g}_n(\hat{h}_n) - g_0(h_0)\|^2 = \frac{1}{n}\sum_{i=1}^n |\hat{g}_n(\hat{h}_n(X_i)) - g_0(h_0(X_i))|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n |\hat{g}_n(\hat{h}_n(X_i)) - g_{0,K}(\hat{h}_n(X_i))|^2 && \mathbf{A} \\
&+ \frac{1}{n}\sum_{i=1}^n |g_{0,K}(\hat{h}_n(X_i)) - g_{0,K}(h_0(X_i))|^2 && \mathbf{B} \\
&+ \frac{1}{n}\sum_{i=1}^n |g_{0,K}(h_0(X_i)) - g_0(h_0(X_i))|^2. && \mathbf{C}
\end{aligned}$$

In the following, we obtain upper bounds for **A**, **B**, and **C**, respectively.

Upper Bound for A: Recall that $\hat{g}_n(\cdot) = B[h(\cdot)]^T\hat{\beta}_n$. By Holder's inequality, **A** can be upper bounded by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n |\hat{g}_n(\hat{h}_n(X_i)) - g_{0,K}(\hat{h}_n(X_i))|^2 \\
&= (\hat{\beta}_n - \beta_{0,K})^T\hat{Q}_{n,K}(\hat{\beta}_n - \beta_{0,K}) \quad \text{where } \hat{Q}_{n,K} = \frac{1}{n}\hat{B}_n^T\hat{B}_n = \frac{1}{n}B(\hat{h}_n)^T B(\hat{h}_n) \\
&\leq \omega_{\max}(\hat{Q}_{n,K})\|\hat{\beta}_n - \beta_{0,K}\|^2. \qquad \text{(B.15)}
\end{aligned}$$

Therefore, it suffices to obtain upper bounds for $\omega_{\max}(\hat{Q}_{n,K})$ and $\|\hat{\beta}_n - \beta_{0,K}\|^2$.

We first obtain an upper bound for $\omega_{\max}(\hat{Q}_{n,K})$. To this end, we need to control the difference between $\hat{Q}_{n,K}$ and Q_K under the spectral norm. First, by Lemma 1.1 of the supplementary of [40],

$$\|\hat{Q}_{n,K} - Q_K\| = O_p[(\xi_{1,K}\delta^*_{h,n})^2 + (\xi_{1,K}\delta^*_{h,n}) + n^{-1/2}\xi_{0,K}(\log K)^{1/2}],$$

which is $o(1)$ because $\xi_{1,K}\delta^*_{h,n} = o(1)$ by Assumption 1.4(ii) and (v), $\xi_{0,K}\sqrt{\frac{\log K}{n}} = o(1)$ by Assumption 1.4(i). Therefore, by Assumption 1.2(iv), we have

$$(2C)^{-1} < \omega_{\min}(\hat{Q}_{n,K}) \leq \omega_{\max}(\hat{Q}_{n,K}) < 2C \tag{B.16}$$

asymptotically.

Next, we obtain an upper bound for $\|\hat{\beta}_n - \beta_{0,K}\|$. For notational convenience, let

$$\begin{aligned}
&\tilde{Y} = [Y_1, \ldots, Y_n]^T,\ G_n = [g_0(h_0(X_1)), \ldots, g_0(h_0(X_n))]^T,\\
&U_n = [u_1, \ldots, u_n]^T,\ \hat{G}_{n,K} = [g_0(\hat{h}_n(X_1)), \ldots, g_0(\hat{h}_n(X_n))]^T,\\
&G_{n,K} = [g_{0,K}(h_0(X_1)), \ldots, g_{0,K}(h_0(X_n))]^T.
\end{aligned}$$

Recall that we consider the model $Y_i = g_0(h(X_i)) + \mu_i$. We have

$$\begin{aligned}
\hat{\beta}_n - \beta_{0,K} &= (\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n^T(\tilde{Y}) - (\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n^T(\hat{B}_n\beta_{0,K})\\
&= (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(\tilde{Y}) - (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T g_{0,K}(\hat{h}_n)\\
&= (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_n + U_n - \hat{G}_{n,K})\\
&= (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_n - G_{n,K}) + (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_{n,K} - \hat{G}_{n,K}) + (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(U_n)
\end{aligned}$$

Therefore, by the triangle inequality, we obtain

$$\begin{aligned}
&\|\hat{\beta}_n - \beta_{0,K}\|^2\\
&= \|(n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_n - G_{n,K}) + (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_{n,K} - \hat{G}_{n,K}) + (n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(U_n)\|^2\\
&\leq \|(n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_n - G_{n,K})\|^2 && \textbf{A.a}\\
&+ \|(n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_{n,K} - \hat{G}_{n,K})\|^2 && \textbf{A.b}\\
&+ \|(n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(U_n)\|^2 && \textbf{A.c}
\end{aligned}$$

It remains to obtain upper bounds for **A.a**, **A.b**, and **A.c**.

Upper Bound for A.a: We have

$$
\begin{aligned}
&\|(n\hat{Q}_{n,K})^{-1}\hat{B}_n^T(G_n - G_{n,K})\|^2 \\
&= n^{-2}(G_n - G_{n,K})^T\hat{B}_n\hat{Q}_{n,K}^{-2}\hat{B}_n(G_n - G_{n,K}) \\
&\leq \omega_{\min}^{-1}(\hat{Q}_{n,K})n^{-2}(G_n - G_{n,K})^T\hat{B}_n\hat{Q}_{n,K}^{-1}\hat{B}_n(G_n - G_{n,K}) \\
&= \omega_{\min}^{-1}(\hat{Q}_{n,K})n^{-1}(G_n - G_{n,K})^T\hat{B}_n(\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n(G_n - G_{n,K}) \\
&= \omega_{\min}^{-1}(\hat{Q}_{n,K})n^{-1}\|\hat{B}_n(\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n(G_n - G_{n,K})\|_2^2 \\
&\text{because } A = \hat{B}_n(\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n \text{ is idempotent with } A = A^2 \\
&\leq \omega_{\min}^{-1}(\hat{Q}_{n,K})n^{-1}\|\hat{B}_n(\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n\|_2^2 \cdot \|G_n - G_{n,K}\|_2^2 \\
&\text{by the Cauchy-Schwarz inequality} \\
&\leq Cn^{-1}\sum_{i=1}^{n}\left[|g_0(h_0(X_i)) - g_{0,K}(h_0(X_i))|^2\right] \\
&= O_p(K^{-2\rho_g}) \quad \text{by Assumption 1.2 (ii),}
\end{aligned}
$$

where the last inequality holds because $\|\hat{B}_n(\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n\|_2^2$, which is the largest eigenvalue of a idempotent matrix, is equal to one (since the eigenvalues of idempotent matrixes are one or zero).

Upper Bound for A.b: Using the same argument as the above, we have

$$
\begin{aligned}
&\|(n\hat{G}_{n,K})^{-1}\hat{B}_n^T(G_{n,K} - \hat{G}_{n,K})\|^2 \\
&= n^{-2}(G_{n,K} - \hat{G}_{n,K})^T\hat{B}_n\hat{Q}_{n,K}^{-2}\hat{B}_n(G_{n,K} - \hat{G}_{n,K}) \\
&\leq \omega_{\min}^{-1}(\hat{Q}_{n,K})n^{-2}(G_{n,K} - \hat{G}_{n,K})^T\hat{B}_n\hat{Q}_{n,K}^{-1}\hat{B}_n(G_{n,K} - \hat{G}_{n,K}) \\
&= \omega_{\min}^{-1}(\hat{Q}_{n,K})n^{-1}(G_{n,K} - \hat{G}_{n,K})^T\hat{B}_n(\hat{B}_n^T\hat{B}_n)^{-1}\hat{B}_n(G_{n,K} - \hat{G}_{n,K}) \\
&\leq Cn^{-1}\sum_{i=1}^{n}\left[|g_{0,K}(h_0(X_i)) - g_{0,K}(\hat{h}_n(X_i))|^2\right].
\end{aligned}
$$

Let $\tilde{h} \in [h_0, \hat{h}_n]$. By the mean value theorem and the Cauchy Schwarz inequality, we obtain

$$
\begin{aligned}
& Cn^{-1}\sum_{i=1}^{n}\Big[|g_{0,K}(h_0(X_i)) - g_{0,K}(\hat{h}_n(X_i))|^2\Big] \\
&= Cn^{-1}\sum_{i=1}^{n}\Big[|\Big(\partial B[\tilde{h}(X_i)]\Big)^T\beta_{0,K}[\hat{h}_n(X_i) - h_0(X_i)]|^2\Big] \\
&\leq Cn^{-1}\sum_{i=1}^{n}\Big\{\Big[\max_{i\leq n}|\Big(\partial B[\tilde{h}(X_i)]\Big)^T\beta_{0,K}|\Big]^2 \cdot |\hat{h}_n(X_i) - h_0(X_i)|^2\Big\} \\
&= C\Big[\max_{i\leq n}|\Big(\partial B[\tilde{h}(X_i)]\Big)^T\beta_{0,K}|\Big]^2 \cdot n^{-1}\sum_{i=1}^{n}\Big\{|\hat{h}_n(X_i) - h_0(X_i)|^2\Big\} \\
&= O_p(\nu_{1,K}^2\delta_{h,n}^{*2}) \quad \text{by Lemma B.1 and definition of } \nu_{1,K}
\end{aligned}
$$

Upper Bound for A.c: First, we have

$$\|(n\hat{Q}_{n,K})^{-1}\hat{B}_n^T U_n\|^2 = U_n^T\hat{B}_n(\hat{Q}_{n,K})^{-2}\hat{B}_n^T U_n n^{-2}.$$

To obtain an upper bound for the above, we use the Markov's inequality. To this end, we need an upper bound of the expected value of the above. First, note that $E[U_nU_n^T \mid \{X_i,\epsilon_i\}_{i=1}^n]$ is a diagonal matrix with diagonal elements equal $E[u_i^2 \mid X_i,\epsilon_i]$. We have

$$
\begin{aligned}
& E[n^{-2}U_n^T\hat{B}_n(\hat{Q}_{n,K})^{-1}\hat{B}_n^T U_n\Big|\{X_i,\epsilon_i\}_{i=1}^n] \\
&= tr\Big\{n^{-2}\hat{B}_n(\hat{Q}_{n,K})^{-1}\hat{B}_n^T E[U_nU_n^T\Big|\{X_i,\epsilon_i\}_{i=1}^n]\Big\} \\
&\leq tr\Big\{n^{-1}(n\hat{Q}_{n,K})^{-1}\hat{B}_n^T\hat{B}_n\cdot C\Big\} \quad \text{by Assumption 1.2(ii)} \\
&= tr(n^{-1}\cdot I_{K\times K}\cdot C) \\
&= O_p(Kn^{-1}),
\end{aligned}
$$

where the second equality holds by the definition of $\hat{Q}_{n,K}$. Therefore by the Markov's Inequality,

$$U_n^T\hat{B}_n(\hat{Q}_{n,K})^{-2}\hat{B}_n U_n n^{-2} = O_p(Kn^{-1}).$$

Combining the upper bounds for **A.a**, **A.b**, and **A.c**, we have

$$\|\hat{\beta}_n - \beta_{0,K}\|^2 = O_p(K^{-2\rho_g} + \nu_{1,K}^2\delta_{h,n}^{*2} + Kn^{-1}).$$

Thus, we have shown that $\mathbf{A} = O_p(K^{-2\rho_g} + \nu_{1,K}^2\delta_{h,n}^{*2} + Kn^{-1})$.

Upper Bound for B: Similar to the upper bound for A.b, by the mean value theorem and Cauchy Schwarz inequality, we have

$$\begin{aligned}
&|g_{0,K}(\hat{h}_n(X_i)) - g_{0,K}(h_0(X_i))|^2 \\
&= \left| \left(\partial B[\tilde{h}(X_i)] \right)^T \beta_{0,K} [\hat{h}(X_i) - h_0(X_i)] \right|^2 \\
&\leq \max_{i \leq n} \left| \left(\partial B[\tilde{h}(X_i)] \right)^T \beta_{0,K} \right|^2 \cdot \left| \hat{h}(X_i) - h_0(X_i) \right|^2,
\end{aligned}$$

where $\tilde{h} \in h_\eta$ lies between $\hat{h}$ and h_0. Thus,

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n |g_{0,K}(\hat{h}_n(X_i)) - g_{0,K}(h_0(X_i))|^2 \leq \max_{i \leq n} \left| \left(\partial B[\tilde{h}(X_i)] \right)^T \beta_{0,K} \right|^2 \cdot \left\{ \frac{1}{n} \sum_{i=1}^n \left| \hat{h}(X_i) - h_0(X_i) \right|^2 \right\} \\
&= O_p(\nu_{1,K}^2) \cdot \|\hat{h} - h_0\|^2 \quad \text{by definition of } \nu_{1,K} \\
&= O_p(\nu_{1,K}^2 \delta_{h,n}^{*2}) \quad \text{by Lemma B.1}
\end{aligned} \tag{B.17}$$

Upper Bound for C: By Assumption 1.2(ii), we have

$$\frac{1}{n} \sum_{i=1}^n |g_{0,K}(h_0(X_i)) - g_0(h_0(X_i))|^2 = O(K^{-2\rho_g}) \tag{B.18}$$

Combining upper bounds for **A**, **B**, and **C**, we have

$$\|\hat{g}_n(\hat{h}_n) - g_0(h_0)\| = O_p(K^{-\rho_g} + \nu_{1,K} \delta_{h,n}^* + K^{1/2} n^{-1/2}) = O_p(\delta_{g,n}^*).$$

□

Lemma B.3. *The following conditions hold:*

(i) Bounded asymptotic variance: $\|v_n^*\|_{sd}$ *satisfies*

$$\liminf_n \|v_n^*\|_{sd} > 0$$

(ii) There exists $g_n \in \mathcal{G}$ *such that* $\|g_n - g_0\|_{\mathcal{G}} = O(\delta_{g,n}^*)$*, and for any* $v_h \in V_1$ *and* $v_g \in V_2$*,* $\|v_h\|_\varphi \leq C_\varphi \|v_h\|_{\mathcal{H}}$ *and* $\|v_g\|_\psi \leq C_\psi \|v_g\|_{\mathcal{G}}$*, where* C_φ *and* C_ψ *are some generic finite positive constant.*

Proof. **Proof of (i):** We have

$$\|v_n^*\|_{sd}^2 = \|g_0(h_0(X)) - \rho[g_0(h_0)]\|^2 + \|d^*_{\Gamma_n}(X)[T - h_0(X)]\|^2 + \|d^*_{g_n}(h_0(X))[Y - g_0(h_0(X))]\|^2$$
$$\geq \|d^*_{g_n}(h_0(X))[Y - g_0(h_0(X))]\|^2$$

By Assumption 1.3(i),

$$\|d^*_{g_n}\| \geq C \quad \text{for all } n.$$

By Assumption 1.2(i),

$$E[u^2 \Big| h_0] \geq C^{-1}\|d^*_{g_n}\|^2 > C. \tag{B.19}$$

So $\liminf_n \|v_n^*\|_{sd} > 0$.

Proof of (ii):

(ii.1) Let $g_{0,K} = B(\cdot)^T\beta_{0,K}$ be g_n, then by Assumption 1.2(iii)

$$\|g_n - g_0\| \leq |g_n - g_0|_d = O(K^{-\rho_g}).$$

Now because $\delta^*_{g,n} = n^{\frac{-1}{2}}K^{\frac{1}{2}} + K^{-\rho_g} + \nu_{1,K}\delta^*_{h,n} \geq K^{-\rho_g}$, we have

$$\|g_n - g_0\| = O(\delta^*_{g,n}).$$

(ii.2) By definition of φ and ψ, we know that

$$< v_1, v_2 >_\varphi = E[v_1 v_2].$$

Let $C_\varphi = C_\psi = 1$, then we have $\|v_h\|_\varphi \leq C_\varphi\|v_h\|_{\mathcal{H}}$ and $\|v_g\|_\psi \leq C_\psi\|v_g\|_{\mathcal{G}}$.

□

Lemma B.4. *(i) Stochastic equicontinuity: for a sequence $k_n = o(n^{-1/2})$, define $g^* = g + k_n u^*_{g_n}$, where $u^*_{g_n} = \frac{d^*_{g_n}(h(x))}{\|v_n^*\|_{sd}}$ then*

$$\sup_{(g,h)\in\mathcal{N}_n} \left|\mu_n\Big\{\psi(o, g^*, h) - \psi(o, g, h) - \triangle_\psi(O, g, h)[k_n u^*_{g_n}]\Big\}\right| = O_p(k_n^2)$$

(ii)

$$\sup_{(g,h)\in\mathcal{N}_n}\left|\mu_n\left\{\triangle_\psi(O,g,h)[u^*_{g_n}]-\triangle_\psi(O,g_0,h_0)[u^*_{g_n}]\right\}\right|=O_p(k_n)$$

(iii) Let $K_\psi(g,h)\equiv E[\psi(o,g,h)-\psi(o,g_0,h_0)]$ *be the Kullback-Leibler type of distance, then*

$$K_\psi(g,h)-K_\psi(g^*,h)=-k_n\Gamma(\alpha_0)[h-h_0,u^*_{g_n}]+\frac{\|g^*-g_0\|^2_\psi-\|g-g_0\|^2_\psi}{2}+O(k_n^2)$$

uniformly over $\alpha=(g,h)\in\mathcal{N}_n$

Proof. We prove (i)-(iii) as follows:

Proof of (i): We first simplify the expression $\psi(o,g,h)-\psi(o,g,h)-\triangle_\psi(O,g,h)[k_nu^*_{g_n}]$. We have

$$\begin{aligned}
&\psi(o,g^*,h)-\psi(o,g,h)-\triangle_\psi(O,g,h)[k_nu^*_{g_n}]\\
&=-\frac{1}{2}\left[\left|Y-g(h(X))-k_nu^*_{g_n}(h(X))\right|^2\right]-\left[-\frac{1}{2}\left|Y-g(h(X))\right|^2\right]-[Y-g(h(X))]\cdot\left[k_nu^*_{g_n}(h(X))\right]\\
&=-\frac{1}{2}k_n^2[u^*_{g_n}(h(X))]^2
\end{aligned}$$

To obtain an upper bound for the last expression, we employ the Markov's inequality. To this end, we need to obtain an upper bound of the expected value of $[u^*_{g_n}(h(X))]^2$. We have

$$\begin{aligned}
E[u^*_{g_n}(h(X))^2]&=\frac{E[d^*_{g_n}(h(X))^2]}{\|g_0(h_0(X))-\rho[g_0(h_0)]\|_2^2+\|d^*_{\Gamma_n}(X)\epsilon\|_2^2+\|d^*_{g_n}(h(X))u\|_2^2}\\
&\le\frac{E[d^*_{g_n}(h(X))^2]}{\|d^*_{g_n}(h(X))u\|_2^2}\le\frac{E[d^*_{g_n}(h(X))^2]}{\|d^*_{g_n}(h(X))\|_2^2\cdot E[u^2]}\\
&\le\frac{E[d^*_{g_n}(h(X))^2]}{C^{-1}\|d^*_{g_n}(h(X))\|_2^2}\quad\text{by Assumption 1.2(i)}\\
&\le C,
\end{aligned}\tag{B.20}$$

where the first inequality hold since $\|d^*_{\Gamma_n}(X)\epsilon\|_2^2\ge 0$. By the Markov's inequality, we obtain

$$\sup_{(g,h)\in\mathcal{N}_n}\left|\mu_n\Big\{-\frac{1}{2}k_n^2[u^*_{g_n}(h(X))]^2\Big\}\right|=\Big\{\sup_{(g,h)\in\mathcal{N}_n}\frac{1}{n}\sum_{i=1}^n[u^*_{g_n}(h(X_i))]^2\Big\}\frac{k_n^2}{2}=O_p(k_n^2).$$

Proof of (ii): To prove (ii) which considers the second order terms, we need to use tools from empirical process. We define a class of functions

$$\mathcal{F}_n = \Big\{ f : (x, h, g) \to \{g[h_0(x)] - g[h(x)]\} u^*_{g_n}(h_0(x)), g, g_0 \in \mathcal{N}_{g,n}, h, h_0 \in \mathcal{N}_{h,n} \Big\},$$

which is the collection of second order them functions of the linearization of $\psi(o, g, h) = \psi(o, g_0, h_0)$.

Note that $\mathcal{F}_n$ is a parametric class indexed by $\theta = (\gamma_L, \beta_K) \in \Theta$ with $\dim(\Theta) = L + K$.

First: For any $f_1 = f(\cdot, g_1, h_1)$ and any $f_2 = f(\cdot, g_2, h_2)$ in $\mathcal{F}_n$ where $h_1, h_2 \in \mathcal{N}_{h,n}$ and $g_1, g_2 \in \mathcal{N}_{g,n}$, i.e. $h_1 = R(\cdot)^T\gamma_1, h_2 = R(\cdot)^T\gamma_2$ and $g_1 = B(\cdot)^T\beta_1, g_2 = B(\cdot)^T\beta_2$, we have

$$\begin{aligned}
&\Big| f_1 - f_2 \Big| \\
&= \Big| [g_0(h_0) - g_1(h_1)] u^*_{g_n}(h_0) - [g_0(h_0) - g_2(h_2)] u^*_{g_n}(h_0) \Big| \\
&= \Big| [g_1(h_1) - g_2(h_2)] u^*_{g_n}(h_0) \Big| \\
&= \Big| [g_1(h_1) - g_1(h_2) + g_1(h_2) - g_2(h_2)] u^*_{g_n}(h_0) \Big| \\
&\leq \Big| [g_1(h_1) - g_1(h_2)] u^*_{g_n}(h_0) \Big| + \Big| [g_1(h_2) - g_2(h_2)] u^*_{g_n}(h_0) \Big|,
\end{aligned}$$

where the last inequality holds by the triangle inequality. By the mean value theorem, for any $\tilde{h} \in h_\eta$ lying between h_1 and h_2, we have

$$\begin{aligned}
&\Big| [g_1(h_1) - g_1(h_2)] u^*_{g_n}(h_0) \Big| + \Big| [g_1(h_2) - g_2(h_2)] u^*_{g_n}(h_0) \Big| \\
&= \Big| \Big(\partial B[\tilde{h}(X_i)] \Big)^T (h_1 - h_2) \beta_1 u^*_{g_n}(h_0) \Big| + \Big| B(h_2)^T (\beta_1 - \beta_2) u^*_{g_n}(h_0) \Big| \\
&\leq \Big| \Big(\partial B[\tilde{h}(X_i)] \Big)^T \beta_1 (h_1 - h_2) u^*_{g_n}(h_0) \Big| + \Big| u^*_{g_n}(h_0) \Big| \cdot \|B(h_2)\| \cdot \|\beta_1 - \beta_2\| \Big| \\
&= \Big| \overset{1\times K}{\Big(\partial B[\tilde{h}(X_i)] \Big)^T} \cdot \overset{K\times 1}{\beta_1} \cdot \overset{1\times K}{R(x)^T} \cdot \overset{K\times 1}{(\gamma_1 - \gamma_2)} \cdot u^*_{g_n}(h_0) \Big| + \Big| u^*_{g_n}(h_0) \Big| \cdot \|B(h_2)\| \cdot \|\beta_1 - \beta_2\| \Big| \\
&= \Big| \Big(\partial B[\tilde{h}(X_i)] \Big)^T (\beta_1 - \beta_{0,K} + \beta_{0,K}) R(x)^T (\gamma_1 - \gamma_2) u^*_{g_n}(h_0) \Big| + \Big| u^*_{g_n}(h_0) \Big| \cdot \|B(h_2)\| \cdot \|\beta_1 - \beta_2\| \Big|.
\end{aligned}$$

Recall from Assumption 1.2(v) that $\sup_{h\in h_\eta}\|B(h)\| \le \xi_{0,K}$. Thus, by the triangle inequality, the last expression can be upper bounded by

$$\begin{aligned}
&\Big\{\|\Big(\partial B[\tilde{h}(X_i)]\Big)^T\| \cdot \|(\beta_1-\beta_{0,K})\| + \Big|\Big(\partial B[\tilde{h}(X_i)]\Big)^T\beta_{0,K}\Big|\Big\}\Big|R(x)^T(\gamma_1-\gamma_2)\Big| \cdot \Big|u^*_{g_n}(h_0)\Big| \\
&+ \xi_{0,K}\cdot\Big|u^*_{g_n}(h_0)\Big|\cdot\|\beta_1-\beta_2\| \\
&\le \Big\{\xi_{1,K}\delta_{2,n}+\nu_{1,K}\Big\}\|R(x)^T\|\cdot\|\gamma_1-\gamma_2\|\cdot\Big|u^*_{g_n}(h_0)\Big| + \xi_{0,K}\cdot\Big|u^*_{g_n}(h_0)\Big|\cdot\|\beta_1-\beta_2\| \\
&\le \Big\{\xi_{1,K}\delta_{2,n}+\nu_{1,K}\Big\}\zeta_{0,L}\cdot\|\gamma_1-\gamma_2\|\cdot\Big|u^*_{g_n}(h_0)\Big| + \xi_{0,K}\cdot\Big|u^*_{g_n}(h_0)\Big|\cdot\|\beta_1-\beta_2\| \\
&\le \Big\{(\xi_{1,K}\delta_{2,n}+\nu_{1,K})\zeta_{0,L}+\xi_{0,K}\Big\}\cdot\Big(\|\gamma_1-\gamma_2\|+\|\beta_1-\beta_2\|\Big)\cdot\Big|u^*_{g_n}(h_0)\Big|,
\end{aligned}$$

where the first inequality holds by Assumption 1.2(v), the definition of $\nu_{1,K}$, that that $\|\beta_1 - \beta_{0,K}\| \le \delta_{2,n}$ since $g_1, g_{0,K} \in \mathcal{N}_{g,n}$, and the second inequality holds by the definition of $\zeta_{0,L}$ in Assumption 1.1(v).

Define $F_n(h_0) = C[(\xi_{1,K}\delta_{2,n}+\nu_{1,K})\zeta_{0,L}+\xi_{0,K}]\cdot\Big|u^*_{g_n}(h_0)\Big|$ then $F_n(h_0)$ is a measurable function such that

$$\Big|f_1 - f_2\Big| \le F_n(h_0)\cdot\Big(\|\gamma_1-\gamma_2\|+\|\beta_1-\beta_2\|\Big)\cdot.$$

Note that $\|F_n(h_0)\|_2 \le C[(\xi_{1,K}\delta_{2,n}+\nu_{1,K})\zeta_{0,L}+\xi_{0,K}]$ because of Eq. (B.20), $E[u^*_{g_n}(h_0)^2] \le C$, and the Markov's inequality.

Second: We obtain the bracketing number of the function class $\mathcal{F}_n$. By Example 19.7 of [100], we have

$$N_{[]}(u\|F_n\|_2, \mathcal{F}_n, \|\cdot\|_2) \le (Cu^{-1})^{L+K},$$

where $u = Y - g_0(h_0(X)) \in (0,1)$, $L+K$ =dim(Θ), and C is a constant.

The bracketing number $N_{[]}(u\|F_n\|_2, \mathcal{F}_n, \|\cdot\|_2)$ goes to infinity as $u \to 0$. A sufficient condition for a class to be Donsker is that the bracketing number does not grow too fast. If the bracketing number for the class $\mathcal{F}_n$ does not grow too fast in the sense that the bracketing integral

$$J_{[]}(d_n, \mathcal{F}_n, L_2(P)) = \int_0^{d_n}[\log(N_{[]}(u\|F_n\|_2, \mathcal{F}_n, \|\cdot\|_2))]^{1/2}du < \infty,$$

where $d_n = \sqrt{E[f^2]}$ as in Eq. B.25 and $f \in \mathcal{F}_n$, then the function class $\mathcal{F}_n$ is a P-Donsker class.

Since $d_n^{-1} \leq Cn$ and $\xi_{F_n} \leq Cn$ as implied by Assumption 1.4, we have

$$J_{[]}(d_n, \mathcal{F}_n, L_2(P)) = \int_0^{d_n} [\log(Cu^{-1})^{L+K}]^{1/2} du < C[(L+K)\log(n)]^{1/2} d_n < \infty. \qquad \text{(B.21)}$$

Therefore $\mathcal{F}_n$ is a P-Donsker class.

Third: In this step, we will apply Lemma 19.36 of [100], which states

For any class $\mathcal{F}_n$ of measurable functions such that for any $f \in \mathcal{F}_n$,

C1 $\|f\|_\infty \leq M_n$

C2 $Pf^2 \leq d_n^2$

then for every f, the empirical process G_n satisfies

1. $E_P[\|G_n\|_{\mathcal{F}_n}] \leq J_{[]}(d_n, \mathcal{F}_n, L_2(P))\Big(1 + J_{[]}(d_n, \mathcal{F}_n, L_2(P))M_n \Big/ (d_n^2\sqrt{n})\Big)$
2. $E_P[\sup_{f\in\mathcal{F}_n} \Big|G_n[f]\Big|] \leq J_{[]}(d_n, \mathcal{F}_n, L_2(P))\Big(1 + J_{[]}(d_n, \mathcal{F}_n, L_2(P))M_n \Big/ (d_n^2\sqrt{n})\Big)$

Thus, we need to verify that the two conditions C1 and C2 are satisfied. To verify condition C1, we have

$$\begin{aligned}
\Big|f(\cdot, g, h)\Big| &= \Big|[g_0(h_0) - g(h)][u^*_{g_n}]\Big| \\
&= \Big|[Y - g(h(X))][u^*_{g_n}] - [Y - g_0(h_0(X))][u^*_{g_n}]\Big| \\
&= \Big|\triangle_\psi(O, g, h)[u^*_{g_n}] - \triangle_\psi(O, g_0, h_0)[u^*_{g_n}]\Big| \\
&= \Big|[g_0(h_0) - g(h_0) + g(h_0) - g(h)][u^*_{g_n}]\Big| \\
&\leq \underbrace{\Big|[g_0(h_0) - g(h_0)\Big|}_{I} \cdot \underbrace{\Big|u^*_{g_n}\Big|}_{II} + \underbrace{\Big|[g(h_0) - g(h)\Big|}_{III} \cdot \Big|u^*_{g_n}\Big|.
\end{aligned}$$

In the following, we obtain upper bounds for I, II, and III, respectively.

Upper Bound for I: By definition, for $g, g_{0,K} \in \mathcal{N}_{g,n}$, $g = B(\cdot)\beta$ and $g_{0,K} = B(\cdot)\beta_{0,K}$, we have

$$
\begin{aligned}
&\Big|[g_0(h_0) - g(h_0)]\Big| = \Big|g_0(h_0) - g_{0,K}(h_0) + g_{0,K}(h_0) - g(h_0)\Big| \\
&\leq \|g_0(h_0) - g_{0,K}(h_0)\|_\infty + \Big|B(h_0)^T(\beta_{0,K} - \beta)\Big| \\
&\leq O(K^{-\rho_g}) + \Big|B(h_0)^T(\beta - \beta_{0,K})\Big| \quad \text{by Assumption 1.2(ii)} \\
&\leq O(K^{-\rho_g}) + \sup_{h \in h_\eta} \|B(h_0)\| \cdot \|\beta - \beta_{0,K}\| \quad \text{by Cauchy-Schwarz Inequality} \\
&\leq O(K^{-\rho_g}) + \xi_{0,K} \cdot \|\beta - \beta_{0,K}\| \quad \text{by definition of } \xi_{0,K} \text{ in Assumption 1.2(v)} \\
&\leq O(K^{-\rho_g}) + \xi_{0,K} \cdot \Big\{\omega_{\min}^{-1}(Q_K)\big(\|g - g_0\|_2 + \|g_0 - g_{0,K}\|_2\big)\Big\} \\
&\leq O(K^{-\rho_g}) + \xi_{0,K} \cdot \Big\{C\big(\|g - g_0\|_2 + \|g_0 - g_{0,K}\|_2\big)\Big\} \quad \text{by Assumption 1.2(iv)} \\
&\leq O(K^{-\rho_g}) + C \cdot \delta_{2,n} \cdot \xi_{0,K} \\
&\leq C\xi_{0,K}\delta_{2,n},
\end{aligned}
$$

where the last inequality holds because for $g, g_{0,K} \in \mathcal{N}_{g,n}$, $\|g - g_0\|_2 + \|g_0 - g_{0,K}\|_2 \leq \delta_{2,n} = \delta^*_{g,n} \log[\log(n)]$ by definition of $\mathcal{N}_{g,n}$, with $\delta^*_{g,n} = K^{\frac{1}{2}} n^{\frac{-1}{2}} + K^{-\rho_g} + \nu_{1,K}\delta^*_{h,n}$ and $\delta^*_{h,n} = L^{\frac{1}{2}} n^{\frac{-1}{2}} + L^{-\rho_h}$. This also indicates that

$$\|\beta - \beta_{0,K}\| \leq C \cdot \delta_{2,n}. \tag{B.22}$$

Upper Bound for II: Recall that $g(\cdot) = B(\cdot)^T\beta_K \in \mathcal{N}_{g,n}$ and let $\tilde{h} \in h_\eta$ which lies between h and h_0. By the mean value theorem, we have

$$
\begin{aligned}
&\Big|g(h_0) - g(h)\Big| \\
&= \Big|[B(h_0)^T - B(h)^T] \cdot \beta\Big| \\
&= \Big|\Big(\partial B[\tilde{h}(X_i)]\Big)^T \cdot [h_0 - h]\beta\Big| \\
&\leq \Big|\Big(\partial B[\tilde{h}(X_i)]\Big)^T \cdot [h_0 - h](\beta - \beta_{0,K})\Big| + \Big|\Big(\partial B[\tilde{h}(X_i)]\Big)^T \cdot [h_0 - h]\beta_{0,K}\Big| \\
&\leq \Big\{\|\frac{\partial B(h(X_i))}{\partial h}\big|_{h=\tilde{h}}\| \cdot \|\beta - \beta_{0,K}\| + \Big|\Big(\partial B[\tilde{h}(X_i)]\Big)^T\beta_{0,K}\Big|\Big\} \cdot \|h - h_0\|_\infty \\
&\leq \Big\{\xi_{1,K} \cdot \|\beta - \beta_{0,K}\| + \sup_{h\in h_\eta}\|\partial B(h)^T\beta_{0,K}\|\Big\} \cdot \|h - h_0\|_\infty \quad \text{by definition of } \xi_{1,K} \text{ in Assumption 1.2(v)} \\
&\leq \Big\{\xi_{1,K} \cdot \|\beta - \beta_{0,K}\| + \nu_{1,K}\Big\} \cdot \|h - h_0\|_\infty \quad \text{by definition of } \nu_{1,K} \\
&\leq \Big\{\xi_{1,K}\delta_{2,n} + \nu_{1,K}\Big\} \cdot \|h - h_0\|_\infty \quad \text{by Eq. (B.22)},
\end{aligned}
$$

where the first inequality holds by the triangle inequality and the second inequality holds by Holder's inequality.

It remains to obtain an upper bound for $\|h - h_0\|_\infty$. Recall that $h = R(\cdot)^T\gamma_L \in \mathcal{N}_{h,n}$. By the triangle inequality, we obtain

$$
\begin{aligned}
&\|h - h_0\|_\infty \leq \|h - h_{0,L}\|_\infty + \|h_{0,L} - h_0\|_\infty \\
&= \|R(x)^T(\gamma_L - \gamma_{0,L})\|_\infty + \|h_{0,L} - h_0\|_\infty \\
&\leq \zeta_{0,L}\|\gamma_L - \gamma_{0,L}\|_\infty + \|h_{0,L} - h_0\|_\infty \quad \text{by definition of } \zeta_{0,L} \\
&\leq \zeta_{0,L}\|\gamma_L - \gamma_{0,L}\|_\infty + O(L^{-\rho_h}) \quad \text{by Assumption 1.1(iii)} \\
&\leq \zeta_{0,L}\Big\{\omega_{\min}^{-1/2}(Q_L)[(\gamma_L - \gamma_{0,L})^T Q_L(\gamma_L - \gamma_{0,L})]^{-1/2}\Big\} + O(L^{-\rho_h}) \\
&= \zeta_{0,L}\omega_{\min}^{-1/2}(Q_L)\|h - h_{0,L}\|_2 + O(L^{-\rho_h}) \\
&\leq \zeta_{0,L}\omega_{\min}^{-1/2}(Q_L)\left[\|h - h_0\|_2 + \|h_0 - h_{0,L}\|_2\right] + O(L^{-\rho_h}) \quad \text{by triangular inequality} \\
&\leq \zeta_{0,L}C\left[\|h - h_0\|_2 + \|h_0 - h_{0,L}\|_2\right] + O(L^{-\rho_h}) \quad \text{by Assumption 1.1(iv)} \\
&\leq \zeta_{0,L}C\delta_{1,n} + O(L^{-\rho_h}),
\end{aligned}
$$

where the last inequality holds because for $h, h_{0,L} \in \mathcal{N}_{h,n}$, $\|h - h_0\|_2 + \|h_0 - h_{0,L}\|_2 \leq \delta_{1,n} = \delta^*_{h,n} \log[\log(n)]$ by definition of $\mathcal{N}_{h,n}$, with $\delta^*_{h,n} = L^{1/2}n^{-1/2} + L^{-\rho_h}$.

Therefore

$$\Big|g(h_0) - g(h)\Big| \leq \Big\{\xi_{1,K}\delta_{2,n} + \nu_{1,K}\Big\} \cdot \Big\{\zeta_{0,L}C\delta_{1,n} + O(L^{-\rho_h})\Big\} \leq C\zeta_{0,L}\delta_{2,n}$$

Upper Bound for III: By definition of $u^*_{g_n}(h_0)$

$$\begin{aligned}
&\sup_{h\in h_\eta}\Big|u^*_{g_n}(h_0)^2\Big| = \sup_{h\in h_\eta}\Big|\overset{1\times K}{\Big(\partial\rho_g(g_0)[P]\Big)^T}\overset{K\times K}{Q_K^{-1}}\overset{K\times 1}{B(h_0)}\overset{1\times K}{B(h_0)^T}\overset{K\times K}{Q_K^{-1}}\overset{K\times 1}{\Big(\partial\rho_g(g_0)[P]\Big)}\Big|\Big/\|v^*_n\|_{sd} \\
&\leq \xi^2_{0,K}\Big|\Big(\partial\rho_g(g_0)[P]\Big)^T Q_k^{-2}\Big(\partial\rho_g(g_0)[P]\Big)\Big|/\|v^*_n\|_{sd} \quad \text{by definition of } \xi_{0,K} \text{ in Assumption 1.2(v)} \\
&\leq \xi^2_{0,K}\Big|\Big(\partial\rho_g(g_0)[P]\Big)^T Q_k^{-2}\Big(\partial\rho_g(g_0)[P]\Big)\Big|/\left[C^{-1}\|d^*_{g_n}\|_2^2\right] \quad \text{by Eq. (B.19) in Lemma B.3} \\
&= \xi^2_{0,K}\Big|\Big(\partial\rho_g(g_0)[P]\Big)^T Q_k^{-2}\Big(\partial\rho_g(g_0)[P]\Big)\Big| \\
&/\left[C^{-1}\|\Big(\partial\rho_g(g_0)[P]\Big)^T Q_k^{-1}B(h_0)B(h_0)^T Q_k^{-1}\Big(\partial\rho_g(g_0)[P]\Big)\|_2^2\right] \\
&\leq C\xi^2_{0,K} \qquad \text{(B.23)}
\end{aligned}$$

Combining the upper bounds for I, II, and III, we have

$$\begin{aligned}
\sup_{f\in\mathcal{F}_n}\|f\|_\infty &= \sup_{(g,h)\in\mathcal{N}_n}\|f\|_\infty \\
&= \sup_{(g,h)\in\mathcal{N}_n}[g_0(h_0) - g(h)]\cdot u^*_{g_n}(h_0) \\
&\leq \sup_{(g,h)\in\mathcal{N}_n}\Big\{\Big|g_0(h_0) - g(h_0)\Big| + \Big|g(h_0) - g(h)\Big|\Big\}\cdot\sup_{h_0\in h_\eta}\Big|u^*_{g_n}(h_0)\Big| \\
&\leq (C\delta_{2,n}\xi_{0,K} + C\zeta_{0,L}\delta_{2,n})C\xi_{0,K} \\
&= C\delta_{2,n}\xi_{0,K}(\xi_{0,K} + \zeta_{0,L}) \equiv M_n \qquad \text{(B.24)}
\end{aligned}$$

Condition C2: We now verify condition C2. We have

$$
\begin{aligned}
E[f^2] &= E\Big\{[g_0(h_0) - g(h)]u^*_{g_n}(h_0)\Big\}^2 \\
&= E\Big\{[g_0(h_0) - g(h_0) + g(h_0) - g(h)]u^*_{g_n}(h_0)\Big\}^2 \\
&\leq 2E\Big\{[g_0(h_0) - g(h_0)]^2 u^*_{g_n}(h_0)^2\Big\} + 2E\Big\{[g(h_0) - g(h)]^2 u^*_{g_n}(h_0)^2\Big\} \\
&\leq C(\xi^2_{0,K}\delta^2_{2,n} + \zeta^2_{0,L}\delta^2_{2,n})E[u^*_{g_n}(h_0)^2] \\
&\leq C\delta^2_{2,n}(\xi^2_{0,K} + \zeta^2_{0,L}) \equiv d_n^2, \qquad \text{(B.25)}
\end{aligned}
$$

where the first inequality holds by using the identity $2xy \leq x^2 + y^2$, and the last inequality holds by Eq. (B.20).

We have verified both conditions C1 and C2. Therefore, by Lemma 19.36 of [100], we obtain

$$
\begin{aligned}
&E[\sup_{(g,h)\in\mathcal{N}_n} \Big|\mu_n\Big\{\triangle_\psi(O,g,h)[u^*_{g_n}(h_0)] - \triangle_\psi(O,g_0,h_0)[u^*_{g_n}(h_0)]\Big\}\Big|] \\
&\leq \frac{J_{[]}(d_n,\mathcal{F}_n,L_2(P))}{\sqrt{n}}(1 + \frac{J_{[]}(d_n,\mathcal{F}_n,L_2(P))}{d_n^2\sqrt{n}} \cdot M_n) \\
&\leq \frac{C[(L+K)\log(n)]^{1/2}d_n}{\sqrt{n}}(1 + \frac{C[(L+K)\log(n)]^{1/2}d_n}{d_n^2\sqrt{n}} \cdot M_n) \text{ by Eq. (B.21)} \\
&= \frac{C[(L+K)\log(n)]^{1/2}d_n}{\sqrt{n}}(1 + \frac{C[(L+K)\log(n)]^{1/2}}{\sqrt{n}} \cdot \frac{M_n}{d_n}) \\
&\leq \frac{C[(L+K)\log(n)]^{1/2}d_n}{\sqrt{n}}(1 + \frac{C[(L+K)\log(n)]^{1/2}}{\sqrt{n}} \cdot \xi_{0,K}) \text{ by Eq. (B.24) and Eq. (B.25)} \\
&= o(1) \quad \text{by Assumptions 1.4(i) and 1.4(v).}
\end{aligned}
$$

By the Markov's inequality, we obtain

$$
\sup_{(g,h)\in\mathcal{N}_n} \Big|\mu_n\left\{\triangle_\psi(O,g,h)[u^*_{g_n}] - \triangle_\psi(O,g_0,h_0)[u^*_{g_n}]\right\}\Big| = o_p(n^{-1/2}).
$$

Proof of (iii): We now obtain an upper bound for the Kullback-Leibler distance. We have

$$
\begin{aligned}
&K_\psi(g,h) - K_\psi(g^*,h) \\
&= E[-\frac{1}{2}\Big|Y - g(h(X))\Big|^2] - E[-\frac{1}{2}\Big|Y - g^*(h(X))\Big|^2] \\
&= E[-\frac{1}{2}\Big|Y - g(h(X))\Big|^2] - E[-\frac{1}{2}\Big|Y - g(h(X)) - \kappa_n u^*_{g_n}(h_0(X))\Big|^2] \\
&= E\Big\{\frac{1}{2}\kappa_n^2[u^*_{g_n}(h_0(X))]^2 - [\kappa_n u^*_{g_n}(h_0(X))][Y - g(h(X))]\Big\} \\
&= E\Big\{\frac{1}{2}\kappa_n^2[u^*_{g_n}(h_0(X))]^2 - [\kappa_n u^*_{g_n}(h_0(X))][Y - g_0(h_0(X))] - [\kappa_n u^*_{g_n}(h_0(X))][g_0(h_0(X)) - g(h(X))]\Big\} \\
&= E\Big\{\frac{1}{2}\kappa_n^2[u^*_{g_n}(h_0(X))]^2 + \kappa_n u^*_{g_n}(h_0(X))][g_0(h_0(X)) - g(h(X))]\Big\} \\
&\text{because } E\{[-\kappa_n u^*_{g_n}(h_0(X))][Y - g_0(h_0(X))]\} = EE\{[-\kappa_n u^*_{g_n}(h_0(X))][Y - g_0(h_0(X))]\Big|h_0(X)\} = 0 \\
&= O(\kappa_n^2) + \kappa_n E\{u^*_{g_n}(h_0(X))[g(h(X)) - g_0(h_0(X))]\} \quad \text{by Eq. (B.20)} \\
&= \kappa_n E\{u^*_{g_n}(h_0(X))[g(h(X)) - g_0(h(X))]\} \\
&+ \kappa_n E\{u^*_{g_n}(h_0(X))[g_0(h(X)) - g_0(h_0(X))]\} + O(\kappa_n^2) \\
&= \kappa_n E\Big\{u^*_{g_n}(h_0(X))\Big[\partial g_0(h_0)(h(X) - h_0(X)) + \partial^2 g_0(\tilde{h})(h(X) - h_0(X))^2\Big]\Big\} \\
&+ \kappa_n E[u^*_{g_n}(h_0(X))\frac{2g(h(X)) + \kappa_n u^*_{g_n}(h_0(X)) - 2g_0(h(X))}{2}] - E[\frac{\kappa_n^2[u^*_{g_n}(h_0(X))]^2}{2}] + O(\kappa_n^2) \\
&\text{by the mean value theorem, where } \tilde{h} \in h_\eta \text{ lies between } h \text{ and } h_0 \\
&= -\kappa_n E\Big\{\partial g_0(h_0)[h(X) - h_0(X)] \cdot [u^*_{g_n}(h_0(X))]\Big\} + \kappa_n E[\partial^2 g_0(\tilde{h})(h(X) - h_0(X))^2 \cdot [u^*_{g_n}(h_0(X))]] \\
&+ E\Big\{\frac{[g^*(h) - g_0(h(X))]^2 - [g(h(X)) - g_0(h(X))]^2}{2}\Big\} - E[\frac{\kappa_n^2[u^*_{g_n}(h_0(X))]^2}{2}] + O(\kappa_n^2) \\
&= -\kappa_n \Gamma(\alpha_0)[h(X) - h_0(X), u^*_{g_n}(h_0(X))] + \kappa_n E[\partial^2 g_0(\tilde{h})(h(X) - h_0(X))^2 \cdot [u^*_{g_n}(h_0(X))]] \\
&+ \frac{[g^*(h) - g_0(h(X))]^2 - [g(h(X)) - g_0(h(X))]^2}{2} + O(\kappa_n^2)
\end{aligned}
$$

by definition of $\Gamma(\alpha_0)[v_h, v_g] = (-1)E\{\partial g_0(h_0)[v_h][v_h]\}$, $\|g\|_\phi^2 = E[g^2]$, and Eq. (B.20).

Therefore, we just need to show that

$$\kappa_n E[\partial^2 g_0(\tilde{h})(h - h_0)^2] = O(\kappa_n^2) \text{ , i.e. } E[\partial^2 g_0(\tilde{h})(h - h_0)^2] = O(\kappa_n).$$

Indeed,

$$
\begin{aligned}
&E[\partial^2 g_0(\tilde{h})(h-h_0)^2] \\
&\leq C \cdot E[(h-h_0)^2] \quad \text{by Assumption 1.2(ii)} \\
&= C \cdot O(\delta_{h,n}^{*2}) \quad \text{by Lemma B.1} \\
&= O(\kappa_n) \quad \text{by Assumption 1.4(iv)-(v)}
\end{aligned}
$$

□

Lemma B.5. *The following holds:*

(i) $\left| < \hat{h}_n - h_0, u_{g_n}^* + u_{\Gamma_n}^* >_\varphi - \mu_n\{\triangle_\varphi(O, h_0)[u_{h_n}^* + u_{\Gamma_n}^*]\} \right| = O_p(\kappa_n)$

(ii) $\frac{1}{n}\sum_{i=1}^n \Big\{ \triangle_\varphi(O, h_0)[u_{h_n}^* + u_{\Gamma_n}^*] + \triangle_\psi(O, g_0, h_0)[u_{g_n}^*] \Big\} \to_d N(0,1)$

(iii) The optimization error in second step $\epsilon_{2,n} = 0$ because the optimization problem has a closed form solution; $\kappa_n(\delta_{2,n}^*)^{-1} = o(1)$; $\|u_{g_n}^*\|_\psi = O(1)$

Proof. We prove (i)-(iii) as follows.

Proof of (i) Because $\rho(\cdot)$ depends on h_0 only through $g_0(h_0)$,

$$
\begin{aligned}
&\left| < \hat{h}_n - h_0, u_{g_n}^* + u_{\Gamma_n}^* >_\varphi - \mu_n\{\triangle_\varphi(O, h_0)[u_{h_n}^* + u_{\Gamma_n}^*]\} \right| \\
&= \Big| < \hat{h}_n - h_{0,L}, u_{g_n}^* + u_{\Gamma_n}^* >_\varphi && \mathbf{D} \\
&+ < h_{0,L} - h_0, u_{g_n}^* + u_{\Gamma_n}^* >_\varphi && \mathbf{E} \\
&- \mu_n\{\triangle_\varphi(O, h_0)[u_{h_n}^* + u_{\Gamma_n}^*]\} \Big| && \mathbf{F}
\end{aligned}
$$

In the following, we simplify **D**, **E**, and **F** by obtaining upper bounds of certain terms.

Simplify D

$$
\begin{aligned}
&< \hat{h}_n - h_{0,L}, u^*_{g_n} + u^*_{\Gamma_n} >_\varphi \\
&= E[u^*_{\Gamma_n}(x) \cdot \overset{1\times L}{R(x)}] \, (\overset{L\times L}{R_n R_n^T}) \overset{L\times n}{R_n} (\overset{n\times 1}{T_n} - H_{n,L}) \\
&= E[u^*_{\Gamma_n}(x) \cdot R(x)^T](R_n R_n^T) R_n (H_n + e_n - H_{n,L}) \\
\text{where } & H_{n,L} = [h_{0,L}(X_1), \ldots, h_0(X_n)]^T \\
& H_n = [h_0(X_1), \ldots, h_0(X_n)]^T, e_n = [\epsilon_1, \ldots, \epsilon_n]^T
\end{aligned}
$$

In the following, we obtain upper bound of $\left| E[u^*_{\Gamma_n}(X) R(X)^T](R_n R_n^T)^{-1} R_n (H_n - H_{n,L}) \right|^2$ to show that $E[u^*_{\Gamma_n}(x) \cdot R(x)^T](R_n R_n^T) R_n (H_n - H_{n,L}) = o_p(\kappa_n)$. By Cauchy-Schwartz inequality,

$$
\begin{aligned}
&\left| E[u^*_{\Gamma_n}(X) R(X)^T](R_n R_n^T)^{-1} R_n (H_n - H_{n,L}) \right|^2 \\
&\leq \| E[u^*_{\Gamma_n}(X) R(X)^T] \|^2 \cdot \| (R_n R_n^T)^{-1} R_n (H_n - H_{n,L}) \|^2 \qquad \text{(B.26)} \\
&= \| E[u^*_{\Gamma_n}(X) R(X)^T] \|^2 \cdot \left[(H_n - H_{n,L})^T R_n^T (R_n R_n^T)^{-2} R_n (H_n - H_{n,L}) \right] \\
&\leq E[u^*_{\Gamma_n}(x)]^2 \cdot E[R(X) R(X)^T] \cdot \left[(H_n - H_{n,L})^T R_n^T (R_n R_n^T)^{-2} R_n (H_n - H_{n,L}) \right] \qquad \text{(B.27)}
\end{aligned}
$$

Note that

$$
\begin{aligned}
\| u^*_{\Gamma_n}(X) \|^2 &= \frac{\| d^*_{\Gamma_n}(X) \|^2}{\| g_0(h_0(X)) - \rho[g_0(h_0)] \|_2^2 + \| d^*_{\Gamma_n}(X) \epsilon \|^2 + \| d^*_{g_n}(x) u \|^2} \\
&\leq \frac{\| d^*_{\Gamma_n}(X) \|^2}{\| d^*_{\Gamma_n}(X) \epsilon \|^2} \leq \frac{\| d^*_{\Gamma_n}(X) \|^2}{\| d^*_{\Gamma_n}(X) \|^2 \cdot E[\epsilon^2]} \\
&\leq \frac{1}{C^{-1}} \quad \text{by Assumption 1.1(i).} \qquad \text{(B.28)}
\end{aligned}
$$

Therefore by Cauchy-Schwartz inequality and Hölder inequality, we know that

$$
E[u^*_{\Gamma_n}(x)]^2 \leq C \qquad \text{(B.29)}
$$

Thus, the above Eq. (B.27) satisfies

$$
\begin{aligned}
& E[u^*_{\Gamma_n}(x)]^2 \cdot E[R(X)R(X)^T] \cdot \left[(H_n - H_{n,L})^T R_n^T (R_n R_n^T)^{-2} R_n (H_n - H_{n,L})\right] && \text{(B.30)} \\
& \leq C \cdot E[R(X)^T R(X)] \cdot \left[(H_n - H_{n,L})^T R_n^T (R_n R_n^T)^{-2} R_n (H_n - H_{n,L})\right] && \text{(B.31)} \\
& \leq C \cdot CL \cdot \left[(H_n - H_{n,L})^T R_n^T (R_n R_n^T)^{-2} R_n (H_n - H_{n,L})\right] \quad \text{by Assumption 1.1(iv)} \\
& = O(CL) \left[(H_n - H_{n,L})^T R_n^T (nQ_{n,L})^{-2} R_n (H_n - H_{n,L})\right] \quad \text{where } Q_{n,L} = \frac{1}{n} R_n R_n^T \\
& = O(CL) \left[(H_n - H_{n,L})^T O(n^{-2}) R_n^T R_n (H_n - H_{n,L})\right] \\
& = O(CL) \left[(H_n - H_{n,L})^T \cdot O(n^{-1}) \cdot (H_n - H_{n,L})\right] \quad \text{by Eq. (B.13)} \\
& = O(CLn^{-1}) \left[(H_n - H_{n,L})^T (H_n - H_{n,L})\right] \\
& = O(CLn^{-1}) O(n \cdot L^{-2\rho_h}) \quad \text{by Assumption 1.1(iii)} \\
& = O(L^{1-2\rho_h}) && \text{(B.32)}
\end{aligned}
$$

By Assumption 1.4(v) we have $O_p(L^{1-2\rho_h}) = o(n^{-1/2}) = o_p(\kappa_n)$. Therefore $E[u^*_{\Gamma_n}(x) \cdot R(x)^T](R_n R_n^T) R_n (H_n - H_{n,L}) = o_p(\kappa_n)$ and

$$
< \hat{h}_n - h_0, u^*_{\Gamma_n} >_\varphi = E[u^*_{\Gamma_n}(X) R(X)^T](R_n R_n)^{-1} R_n e_n + o_p(\kappa_n).
$$

Upper bound for E We obtain upper bound for **E**. Consider

$$
\begin{aligned}
& \left| < h_{0,L} - h_0, u^*_{\Gamma_n} >_\varphi \right| \leq \|h_{0,L} - h_0\|_2 \cdot \|u^*_{\Gamma_n}\|_2 \quad \text{by Hölder inequality} \\
& \leq \|h_{0,L} - h_0\|_2 \cdot \left\{ \frac{1}{C^{-1}} \right\} \quad \text{by Eq. (B.28)} \\
& = O(L^{-\rho_h}) \quad \text{by Assumption 1.1(iii)} \\
& = o(\kappa_n) \quad \text{by Assumption 1.4(v).}
\end{aligned}
$$

F

$$\mu_n\Big\{\triangle_\varphi(O,h_0)[u^*_{\Gamma_n}]\Big\}$$
$$= \mu_n\Big\{E[\partial g_0(h_0)d^*_{g_n}(h_0(X))R(X)^T]Q_L^{-1}R(X)\frac{1}{\|v_n^*\|_{sd}}\cdot[T-h_0(X)]\Big\}$$

by definition of $u^*_{\Gamma_n}$ and $\triangle_\varphi(O,h_0) = T - h_0(X)$

$$= E[\partial g_0(h_0)d^*_{g_n}(h_0(X))R(X)^T]Q_L^{-1}\frac{1}{\|v_n^*\|_{sd}}\cdot\mu_n\Big\{\cdot[T-h_0(X)]R(X)\Big\}$$
$$= E[\partial g_0(h_0)d^*_{g_n}(h_0(X))R(X)^T]Q_L^{-1}\frac{1}{\|v_n^*\|_{sd}}\cdot\frac{1}{n}\sum_{i=1}^n[T_i-h_0(X_i)]R(X_i)$$

because $\dfrac{1}{n}\displaystyle\sum_{i=1}^n[T_i-h_0(X_i)]R(X_i)$ has mean zero

$$= E[\partial g_0(h_0)d^*_{g_n}(h_0(X))R(X)^T]\Big\{Q_L^{-1}\frac{1}{\|v_n^*\|_{sd}}\cdot\frac{1}{n}R_ne_n\Big\}$$
$$= E[\partial g_0(h_0)d^*_{g_n}(h_0(X))R(X)^TQ_L^{-1}R(X)R(X)^T]\Big\{Q_L^{-1}\frac{1}{\|v_n^*\|_{sd}}\cdot\frac{1}{n}R_ne_n\Big\}$$

because $Q_L^{-1}E[R(X)R(X)^T] = 1$

$$= E[\Big\{\partial g_0(h_0)d^*_{g_n}(h_0(X))R(X)^TQ_L^{-1}R(X)\Big\}\cdot R(X)^T]\Big\{Q_L^{-1}\frac{1}{\|v_n^*\|_{sd}}\cdot\frac{1}{n}R_ne_n\Big\}$$
$$= E[d^*_{\Gamma_n}(X)R(X)^T]\Big\{Q_L^{-1}\frac{1}{\|v_n^*\|_{sd}}\cdot\frac{1}{n}R_ne_n\Big\}$$
$$= E[u^*_{\Gamma_n}(X)R(X)^T]Q_L^{-1}\cdot\frac{1}{n}R_ne_n.$$

Combining **D**, **E**, and **F**, we have

D+**E**+**F**

$$= E[u^*_{\Gamma_n}(X)R(X)^T](nQ_{n,L})^{-1}R_ne_n + o_p(\kappa_n) - E[u^*_{\Gamma_n}(X)R(X)^T](nQ_L)^{-1}R_ne_n + o_p(\kappa_n)$$
$$= E[u^*_{\Gamma_n}(X)R(X)^T](Q_{n,L}^{-1}-Q_L^{-1})\frac{R_ne_n}{n} + o_p(\kappa_n). \tag{B.33}$$

Next we obtain upper bound of Eq. (B.33). Consider

$$
\begin{aligned}
&\left|E[u^*_{\Gamma_n}(x)R(X)^T](Q^{-1}_{n,L}-Q^{-1}_L)\frac{R_ne_n}{n}\right|^2 \\
&= \left|E[u^*_{\Gamma_n}(X)R(X)^T]\ (Q^{-1}_{n,L}-Q^{-1}_L)\overset{L\times n}{R_n}e_n\overset{n\times L}{e_n^T}R_n^T(Q^{-1}_{n,L}-Q^{-1}_L)\ E[u^*_{\Gamma_n}R(X)]\right|/n^2 \\
&= \left|\mathrm{tr}\Big\{E[u^*_{\Gamma_n}(X)R(X)^T]\ (Q^{-1}_{n,L}-Q^{-1}_L)\overset{L\times n}{R_n}e_n\overset{n\times L}{e_n^T}R_n^T(Q^{-1}_{n,L}-Q^{-1}_L)\ E[u^*_{\Gamma_n}R(X)]\Big\}\right|/n^2 \\
&= \left|\mathrm{tr}\Big\{E[R_ne_ne_n^TR_n^T(Q^{-1}_{n,L}-Q^{-1}_L)]\cdot E[(u^*_{\Gamma_n})^2R(X)R(X)^T(Q^{-1}_{n,L}-Q^{-1}_L)]\Big\}\right|/n^2 \\
&= \left|C\cdot\mathrm{tr}\Big\{E[R_nR_n^T(Q^{-1}_{n,L}-Q^{-1}_L)]\cdot E[(u^*_{\Gamma_n})^2R(X)R(X)^T(Q^{-1}_{n,L}-Q^{-1}_L)]\Big\}\right|/n^2 \quad \text{by Assumption 1.1(ii)} \\
&= \left|C\cdot\mathrm{tr}\Big\{E[nQ_{n,L}(Q^{-1}_{n,L}-Q^{-1}_L)]\cdot E[(u^*_{\Gamma_n})^2Q_L(Q^{-1}_{n,L}-Q^{-1}_L)]\Big\}\right|/n^2 \quad \text{by definition of } Q\,\mathrm{ln} \text{ and } Q_L \\
&= \left|C\cdot\mathrm{tr}\Big\{E[Q_{n,L}(Q^{-1}_{n,L}-Q^{-1}_L)]\cdot E[C\cdot Q_L(Q^{-1}_{n,L}-Q^{-1}_L)]\Big\}\right|/n \quad \text{by Eq. (B.29)} \\
&= \left|C\cdot\mathrm{tr}\Big\{E[(Q_{n,L}-Q_L)]\cdot(Q^{-1}_{n,L}-Q^{-1}_L)]\Big\}\right|/n \\
&\le \|Q_{n,L}-Q_L\|\cdot[\omega_{\min}(Q_{n,L})\cdot L+\omega_{\min}(Q_L)\cdot L]/n \\
&\le \|Q_{n,L}-Q_L\|\cdot[C\cdot(L+L)]/n \quad \text{by Eq. (B.13) and Assumption 1.1(iv)} \\
&= O(\zeta_{0,L}(\log L)^{\frac{1}{2}}n^{-\frac{1}{2}}\cdot L\cdot n^{-1}) \quad \text{by Eq. (B.12)} \\
&= O(n^{-1}). \quad \text{by Assumption 1.4(iv)}
\end{aligned}
$$

The above indicates that $\left|\mathbf{D}+\mathbf{E}+\mathbf{F}\right|^2 < O_p(n^{-1})$. Therefore

$$\left| < \hat{h}_n - h_0, u^*_{g_n}+u^*_{\Gamma_n} >_\varphi - \mu_n\{\triangle_\varphi(O,h_0)[u^*_{h_n}+u^*_{\Gamma_n}]\}\right| = \mathbf{D}+\mathbf{E}+\mathbf{F} = O_p(\kappa_n)$$

Proof of (ii): By definition,

$$\triangle_\varphi(O,h_0)[u^*_{\Gamma_n}]+\triangle_\psi(O,g_0,h_0)[u^*_{g_n}] = \frac{u^*_{\Gamma_n}(X)\epsilon+u^*_{g_n}(h_0(X))u}{\|v^*_n\|_{sd}}.$$

Now we first consider

$$\sup_{x\in\mathcal{X}} \frac{|d^*_{\Gamma_n}(X)|^2}{\|v^*_n\|_{sd}}$$
$$= \frac{\zeta^2_{0,L}}{\omega|_{\min}(Q_L)} \cdot \frac{\|E[\partial g_0(h_0) d^*_{g_n}(h_0(X))R(X)]\|^2}{\|v^*_n\|_{sd}}$$
$$\leq \frac{C\zeta^2_{0,L}}{\omega|^2_{\min}(Q_L)} \frac{\|E[\partial g_0(h_0) d^*_{g_n}(h_0(X))R(X)]\|^2}{\|d^*_{g_n}\|^2} \quad \text{by Assumption 1.1(iv) and definition of } \|v^*_n\|_{sd}$$
$$\leq \frac{C\zeta^2_{0,L}}{\omega|^2_{\min}(Q_L)} \frac{E\left[\partial g_0(h_0) d^*_{g_n}(h_0(X))\right]^2 E\left[R(X)^T R(X)\right]}{\|d^*_{g_n}\|^2} \quad \text{by Cauchy-Schwartz inequality}$$
$$\leq \frac{C\zeta^2_{0,L}}{\omega|_{\min}(Q_L)} \frac{\sup_{h\in h_\eta}\left[\partial g_0(h_0)\right]^2 E\left[d^*_{g_n}(h_0(X))\right]^2 E\left[R(X)^T R(X)\right]}{\|d^*_{g_n}\|^2}$$
$$\leq O(L\zeta^2_{0,L}) \quad \text{by Assumption 1.1(iv)(v) and Assumption 1.2(ii).} \tag{B.34}$$

Second, we consider

$$\frac{E[\left|d^*_{\Gamma_n}(X)\epsilon + d^*_{g_n}(h_0(X))u\right|]}{n\|v^*_n\|^4_{sd}}$$
$$\leq 8 \cdot \frac{E[\left|d^*_{\Gamma_n}(X)\epsilon\right|^4] + E[\left|d^*_{g_n}(h_0(X))u\right|^4]}{n\|v^*_n\|^4_{sd}}$$
$$\leq C \cdot \frac{E[\left|d^*_{\Gamma_n}(X)\right|^4] + E[\left|d^*_{g_n}(h_0(X))\right|^4]}{n\|v^*_n\|^4_{sd}} \quad \text{by Assumption 1.1(ii) and 1.2(i)}$$
$$\leq C \cdot n^{-1} \cdot (\xi_{0,K} + L\zeta_{0,L})(E[|u^*_{\Gamma_n}|^2] + E[|u^*_{g_n}|^2])$$
by Eq. (B.23) that $\sup_{h\in h_\eta} |u^*_{g_n}|^2 < C\xi^2_{0,K}$, and Eq. (B.34) that $\sup_{x\in\mathcal{X}} |u^*_{\Gamma_n}|^2 < CL\zeta^2_{0,L}$
$$\leq C(n^{-1}\xi^2_{0,K} + n^{-1}L\zeta^2_{0,L})$$
by Eq. (B.20) that $E[|u^*_{g_n}|^2] < C$, and Eq. (B.28) that $E[|u^*_{\Gamma_n}|^2] < C$
$$= o(1) \quad \text{by Assumption 1.4(i).}$$

From the above two conclusions, by Assumption 1.1(i) and the Linderberge Central Limit Theorem,

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\triangle_\varphi(O,h_0)[u^*_{h_n} + u^*_{\Gamma_n}] + \triangle_\psi(O,g_0,h_0)[u^*_{g_n}]\Big\} \to_d N(0,1).$$

Proof of (iii): First, the second step optimization error $\epsilon_{2,n} = o(n^{-1/2})$ because we use direct projection onto spline basis functions.

Second, by Assumption 1.4 (iii), $\kappa_n(\delta^*_{2,n})^{-1} = o(1)$.

Third, $\|u^*_{g_n}\|^2_\psi = E[|u^*_{g_n}|^2] < C$ by Eq. (B.23). Thus $\|u^*_{g_n}\|_\psi \leq E[|u^*_{g_n}|] = O(1)$. □

Proof of Theorem 3.2

Theorem 3.2 Under Assumptions 1.1-1.4, we have

$$\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_0(h_0)] = \frac{1}{n}\sum_{i=1}^{n}\Big\{\triangle_\varphi(O,h_0)[d^*_{\Gamma_n}] + \triangle_\psi(O,g_0,h_0)[d^*_{g_n}]\Big\} + \|v^*_n\|_{sd} o(n^{-1/2}),$$

where

$$\triangle_\varphi(O_i,h_0)[d^*_{\Gamma_n}] = (-1)\frac{E[Y|T=1,X_i] - E[Y|T=1,h_0(X_i)]}{h_0(X_i)}\{T_i - E[Y|T=1,h_0(X_i)]\} + o(n^{-1/2})$$

and

$$\triangle_\psi(O_i,g_0,h_0)[d^*_{g_n}] = \frac{T_i}{h_0(X_i)}\{Y_i - E[Y|T=1,h_0(X_i)]\} + o(n^{-1/2})$$

Proof. We first linearize $\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_{0,n}(h_{0,n})]$, then by Assumption 3.1 we can show that $|\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_0(h_0)]| \le \|v^*_n\|_{sd} \cdot o(n^{-1/2})$.

First of all, consider

$$\begin{aligned}
&\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_{0,n}(h_{0,n})] \\
&= \Big\{\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_0(h_0)]\Big\} - \Big\{\rho[g_{0,n}(h_{0,n})] - \rho[g_0(h_0)]\Big\} \\
&- \Big\{\frac{\partial\rho[g_0(h)]}{\partial h}[\hat{h}_n - h_0] + \frac{\partial\rho[g_0(h)]}{\partial g}[\hat{g}_n - g_0]\Big\} + \Big\{\frac{\partial\rho[g_0(h)]}{\partial h}[h_{0,n} - h_0] + \frac{\partial\rho[g_0(h)]}{\partial g}[g_{0,n} - g_0]\Big\} \\
&+ \Big\{\frac{\partial\rho[g_0(h)]}{\partial h}[\hat{h}_n - h_{0,n}] + \frac{\partial\rho[g_0(h)]}{\partial g}[\hat{g}_n - g_{0,n}]\Big\} \\
&= \Big\{\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_0(h_0)] - \frac{\partial\rho[g_0(h)]}{\partial h}[\hat{h}_n - h_0] + \frac{\partial\rho[g_0(h)]}{\partial g}[\hat{g}_n - g_0]\Big\} && <1> \\
&- \Big\{\rho[g_{0,n}(h_{0,n})] - \rho[g_0(h_0)] - \frac{\partial\rho[g_0(h)]}{\partial h}[h_{0,n} - h_0] + \frac{\partial\rho[g_0(h)]}{\partial g}[g_{0,n} - g_0]\Big\} && <2> \\
&+ \Big\{\frac{\partial\rho[g_0(h)]}{\partial h}[\hat{h}_n - h_{0,n}] + \frac{\partial\rho[g_0(h)]}{\partial g}[\hat{g}_n - g_{0,n}]\Big\}, && <3>
\end{aligned}$$

where

$<1> = o(n^{-1/2})$ by Assumption 3.1 (ii) and the fact that $\hat{\alpha}_n = (\hat{g}_n, \hat{h}_n) \in \mathcal{N}_n$,

$<2> = o(n^{-1/2})$ by Assumption 3.1 (ii) and (iii) which implies that $(g_{0,n}, h_{0,n}) \in \mathcal{N}_n$.

Thus

$$\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_{0,n}(h_{0,n})] = \frac{\partial \rho[g_0(h)]}{\partial h}[\hat{h}_n - h_{0,n}] + \frac{\partial \rho[g_0(h)]}{\partial g}[\hat{g}_n - g_{0,n}] + o(n^{-1/2})$$

By Riesz Representation, there exist $d^*_{h_n} \in V_{1,n}$ and $d^*_{g_n} \in V_{2,n}$ such that

$$\begin{aligned}
&\frac{\partial \rho[g_0(h)]}{\partial h}[\hat{h}_n - h_{0,n}] + \frac{\partial \rho[g_0(h)]}{\partial g}[\hat{g}_n - g_{0,n}] \\
&=< d^*_{h_n}, \hat{h}_n - h_{0,n} >_\varphi + < d^*_{g_n}, \hat{g}_n - g_{0,n} >_\psi \\
&=< d^*_{h_n}, \hat{h}_n - h_0 >_\varphi + < d^*_{h_n}, h_0 - h_{0,n} >_\varphi + < d^*_{g_n}, \hat{g}_n - g_0 >_\psi + < d^*_{g_n}, g_0 - g_{0,n} >_\psi \\
&=< d^*_{h_n}, \hat{h}_n - h_0 >_\varphi + < d^*_{g_n}, \hat{g}_n - g_0 >_\psi
\end{aligned}$$

by the orthogonality that

$$h_{0,n} = \Pi_{V_{1,n}}(h_0), d^*_{h_n} \in V_{1,n}$$

$$g_{0,n} = \Pi_{V_{2,n}}(g_0), d^*_{g_n} \in V_{2,n}.$$

Because ρ depends on h only through $g(h)$, we have

$$\frac{\partial \rho[g_0(h)]}{\partial h}[\hat{h}_n - h_{0,n}] + \frac{\partial \rho[g_0(h)]}{\partial g}[\hat{g}_n - g_{0,n}] =< d^*_{g_n}, \hat{g}_n - g_0 >_\psi .$$

Therefore,

$$\begin{aligned}
&\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_{0,n}(h_{0,n})] \\
&=< d^*_{g_n}, \hat{g}_n - g_0 >_\psi =< d^*_{g_n}, \hat{g}_n(\hat{h}_n) - g_0(h_0) >_\psi \\
&= P_n \triangle_\psi(O, g_0, h_0)[d^*_{g_n}] + < d^*_{\Gamma_n}, \hat{h}_n - h_0 >_\varphi \\
&+ \Big\{ < d^*_{g_n}, \hat{g}_n - g_0 >_\psi - P_n \triangle_\psi(O, g_0, h_0)[d^*_{g_n}] - < d^*_{\Gamma_n}, \hat{h}_n - h_0 >_\varphi \Big\} \\
&= P_n \Big\{ \triangle_\psi(O, g_0, h_0)[d^*_{g_n}] + \triangle_\varphi(O, h_0)[d^*_{\Gamma_n}] \Big\} \\
&+ \Big\{ < d^*_{g_n}, \hat{g}_n - g_0 >_\psi - P_n \triangle_\psi(O, g_0, h_0)[d^*_{g_n}] - < d^*_{\Gamma_n}, \hat{h}_n - h_0 >_\varphi \Big\}
\end{aligned}$$

If we can show that

$$\Big| < d^*_{g_n}, \hat{g}_n - g_0 >_\psi - P_n \triangle_\psi(O, g_0, h_0)[d^*_{g_n}] - < d^*_{\Gamma_n}, \hat{h}_n - h_0 >_\varphi \Big| = \|v^*_n\|_{sd} \cdot o(n^{-1/2}), \quad \text{(B.35)}$$

then we will have

$$\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_{0,n}(h_{0,n})] = P_n\Big\{\triangle_\psi(O, g_0, h_0)[d^*_{g_n}] + \triangle_\varphi(O, h_0)[d^*_{\Gamma_n}]\Big\} + \|v^*_n\|_{sd} \cdot o(n^{-1/2}),$$

which provides a linearization of the estimator that allows us to study the asymptotic distribution of the estimator.

In the following, we are going to show Eq. (B.35).

Recall that we define standardized Riesz representer $(u^*_{g_n}, u^*_{h_n}, u^*_{\Gamma_n}) = \frac{1}{\|v^*_n\|_{sd}}(d^*_{g_n}, d^*_{h_n}, d^*_{\Gamma_n})$. Define $\hat{g}^*_n = \hat{g}_n + \kappa_n u^*_{g_n}$ with a sequence $\kappa_n = o(n^{-1/2})$. Because $\hat{g}_n \in \mathcal{N}_{g,n}$, $\|\hat{g}_n - g_0\|_\psi = O(\delta^*_{2,n})$. In addition, by Assumption 3.3 (iii), $\|u^*_{g_n}\| = O(1)$. Then we have

$$\begin{aligned}
\|\hat{g}^*_n - g_0\|_\psi &\le \|\hat{g}_n - g_0\|_\psi + \|\kappa_n u^*_{g_n}\| \\
&= O_p(\delta^*_{2,n}) + \kappa_n \cdot O(1) \\
&= O_p(\delta^*_{2,n}) \quad \text{by Assumption 3.3 (iii).}
\end{aligned}$$

Thus $\hat{g}^*_n \in \mathcal{N}_{g,n}$ wpa1. By definition of $\hat{g}_n$,

$$P_n\psi(O, \hat{g}_n, \hat{h}_n) \ge \sup_{g \in \mathcal{G}_n} P_n\psi(O, g, \hat{h}_n) - O_p(\epsilon^2_{2,n}) \ge P_n\psi(O, \hat{g}^*_n, \hat{h}_n) - O_p(\epsilon^2_{2,n}).$$

That is

$$\begin{aligned}
&- O_p(\epsilon^2_{2,n}) \\
&\le P_n\psi(O, \hat{g}_n, \hat{h}_n) - P_n\psi(O, \hat{g}^*_n, \hat{h}_n) \\
&= \mu_n\Big\{\psi(O, \hat{g}_n, \hat{h}_n) - \psi(O, \hat{g}^*_n, \hat{h}_n)\Big\} + \Big\{K_\psi(\hat{g}_n, \hat{h}_n) - K_\psi(\hat{g}^*_n, \hat{h}_n)\Big\} \\
&= \mu_n\Big\{\psi(O, \hat{g}_n, \hat{h}_n) - \psi(O, \hat{g}^*_n, \hat{h}_n) + \triangle_\psi(O, \hat{g}_n, \hat{h}_n)[\hat{g}^*_n - \hat{g}_n]\Big\} && < 1 > \\
&+ \mu_n\Big\{\triangle_\psi(O, g_0, h_0)[\hat{g}^*_n - \hat{g}_n] - \triangle_\psi(O, \hat{g}_n, \hat{h}_n)[\hat{g}^*_n - \hat{g}_n]\Big\} && < 2 > \\
&+ \mu_n\Big\{ - \triangle_\psi(O, g_0, h_0)[\hat{g}^*_n - \hat{g}_n]\Big\} + \Big\{K_\psi(\hat{g}_n, \hat{h}_n) - K_\psi(\hat{g}^*_n, \hat{h}_n)\Big\}.
\end{aligned}$$

By Assumption 3.2(i), we have $< 1 >= O_p(k_n^2)$ and $< 2 >= O_p(k_n)$ and thus

$$
\begin{aligned}
&- O_p(\epsilon_{2,n}^2) \\
&\leq P_n\psi(O,\hat{g}_n,\hat{h}_n) - P_n\psi(O,\hat{g}_n^*,\hat{h}_n) \\
&= -\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[\hat{g}_n^* - \hat{g}_n]\Big\} + \Big\{K_\psi(\hat{g}_n,\hat{h}_n) - K_\psi(\hat{g}_n^*,\hat{h}_n)\Big\} + O_p(k_n) \\
&= -\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[\hat{g}_n^* - \hat{g}_n]\Big\} + \Big\{(-k_n)\Gamma(\alpha_0)[\hat{h}_n - h_0, u_{g_n}^*] + \frac{\|\hat{g}_n^* - g_0\|_\psi^2 - \|\hat{g}_n - g_0\|_\psi^2}{2}\Big\} + O_p(k_n) \\
&\text{by Assumption 3.2(ii)} \\
&\leq -\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[k_n u_{g_n}^*]\Big\} + \Big\{(-k_n)\Gamma(\alpha_0)[\hat{h}_n - h_0, u_{g_n}^*] + \frac{\|k_n u_{g_n}^*\|_\psi^2 + 2k_n\|u_{g_n}^*(\hat{g}_n - g_0)\|_\psi}{2}\Big\} + O_p(k_n) \\
&= -\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[k_n u_{g_n}^*]\Big\} + \Big\{(-k_n) < \hat{h}_n - h_0, u_{\Gamma_n}^* >_\varphi + \frac{\|k_n u_{g_n}^*\|_\psi^2 + 2k_n\|u_{g_n}^*(\hat{g}_n - g_0)\|_\psi}{2}\Big\} + O_p(k_n) \\
&\text{by definition of } \Gamma(\alpha_0) \text{ and } u_{\Gamma_n}^* \\
&= -\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[k_n u_{g_n}^*]\Big\} + \Big\{(-k_n) < \hat{h}_n - h_0, u_{\Gamma_n}^* >_\varphi + O(k_n^2) + k_n\|u_{g_n}^*(\hat{g}_n - g_0)\|_\psi\Big\} + O_p(k_n) \\
&\text{because } \|u_{g_n}^*\|_\psi = O(1) \text{ in Assumption 3.3(iii)} \\
&= -\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[k_n u_{g_n}^*]\Big\} + \Big\{(-k_n) < \hat{h}_n - h_0, u_{\Gamma_n}^* >_\varphi + O(k_n^2) + k_n < \hat{g}_n - g_0, u_{g_n}^* >_\psi \Big\} + O_p(k_n) \\
&\text{by definition of } \|\cdot\|_\psi
\end{aligned}
$$

By Assumption 3.3 (iii), $O_p(\epsilon_{2,n}^2) = O_p(k_n^2)$. Therefore,

$$O_p(\epsilon_{2,n}^2) = O_p(k_n^2) \geq \mu_n\Big\{\triangle_\psi(O,g_0,h_0)[k_n u_{g_n}^*]\Big\} + \Big\{-k_n < \hat{g}_n - g_0, u_{g_n}^* >_\psi + k_n < \hat{h}_n - h_0, u_{\Gamma_n}^* >_\varphi \Big\} + O_p(k_n)$$

Similarly, let $\hat{g}_n^* = \hat{g}_n - \kappa_n u_{g_n}^*$ and we have

$$O_p(\epsilon_{2,n}^2) = O_p(k_n^2) \geq -\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[k_n u_{g_n}^*]\Big\} + \Big\{+k_n < \hat{g}_n - g_0, u_{g_n}^* >_\psi - k_n < \hat{h}_n - h_0, u_{\Gamma_n}^* >_\varphi \Big\} + O_p(k_n)$$

In summary, we have

$$O_p(k_n^2) \geq \Big|\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[k_n u_{g_n}^*]\Big\} + \Big\{k_n < \hat{h}_n - h_0, u_{\Gamma_n}^* >_\varphi - k_n < \hat{g}_n - g_0, u_{g_n}^* >_\psi \Big\}\Big|.$$

Dividing both sides by k_n we have

$$O_p(k_n) = \Big|\mu_n\Big\{\triangle_\psi(O,g_0,h_0)[u_{g_n}^*]\Big\} + \Big\{ < \hat{h}_n - h_0, u_{\Gamma_n}^* >_\varphi - < \hat{g}_n - g_0, u_{g_n}^* >_\psi \Big\}\Big|,$$

which is equivalent to Eq. (B.35). Thus we have

$$\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_{0,n}(h_{0,n})] = P_n\Big\{\triangle_\psi(O, g_0, h_0)[d^*_{g_n}] + \triangle_\varphi(O, h_0)[d^*_{\Gamma_n}]\Big\} + \|v^*_n\|_{sd} \cdot o(n^{-1/2}).$$

Now consider

$$\begin{aligned}
&\frac{1}{\|v^*_n\|_{sd}}\Big|\rho[g_{0,n}(h_{0,n})] - \rho[g_0(h_0)]\Big| \\
&\leq \frac{1}{\|v^*_n\|_{sd}}\Big\{\Big|\rho[g_{0,n}(h_{0,n})] - \rho[g_0(h_0)] - \frac{\partial\rho}{\partial h}[h_{0,n} - h_0] - \frac{\partial\rho}{\partial g}[g_{0,n} - g_0]\Big| \\
&+ \Big|\frac{\partial\rho}{\partial h}[h_{0,n} - h_0]\Big| + \Big|\frac{\partial\rho}{\partial g}[g_{0,n} - g_0]\Big|\Big\} \\
&= o(n^{-1/2}) \quad \text{by Assumption 3.1.}
\end{aligned}$$

Therefore

$$\rho[\hat{g}_n(\hat{h}_n)] - \rho[g_0(h_0)] = P_n\Big\{\triangle_\psi(O, g_0, h_0)[d^*_{g_n}] + \triangle_\varphi(O, h_0)[d^*_{\Gamma_n}]\Big\} + \|v^*_n\|_{sd} \cdot o(n^{-1/2}),$$

where

$$\begin{aligned}
\triangle_\varphi(O_i, h_0) &= T_i - h_0(X_i) \\
\triangle_\psi(O_i, g_0, h_0) &= \frac{T_i}{h_0(X_i)}[Y_i - g_0[h_0(X_i)]].
\end{aligned}$$

Now we calculate $d^*_{g_n}(h_0)$ and $d^*_{\Gamma_n}(x)$. By definition,

$$\begin{aligned}
d^*_{g_n}(h_0) &= \partial\rho_g(g_0)[B]^T Q^{-1}_{K(n)} B(h_0) \quad \text{and} \\
d^*_{\Gamma_n}(x) &= E[\partial g_0(h_0) d^*_{g_n}(h_0) R(X)^T] Q^{-1}_{L(n)} R(x).
\end{aligned}$$

First, we can see that $\partial\rho_g(g_0)[B]^T = E[\Big(\int 1 dP(x)\Big)B(h_0)] = E[B(h_0)]$. Second, we consider calculating $E[\partial g_0(h_0)]$. Denote $h_\theta = h_0 + \theta * r(x)$, where $r(x)$ is an element of $R(x) = [r_1(x), \ldots, r_L(x)]^T$. Then we can see that $\frac{\partial}{\partial\theta}\Big|_{\theta=0} h_\theta = r(x)$ and $\frac{\partial}{\partial\theta}\Big|_{\theta=0} \frac{h_0}{h_\theta} = -\frac{r(x)}{h_0}$.

Now consider

$$
\begin{aligned}
& E[\partial g_0(h_0)] \\
&= E[\frac{\partial}{\partial\theta}\Big|_{\theta=0} g_0(h_\theta)] \\
&= E[\Big(\frac{h_0}{h_\theta}\Big|_{\theta=0}\Big)\cdot\Big(\frac{\partial}{\partial\theta}\Big|_{\theta=0} g_0(h_\theta)\Big)] \\
&= E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\Big(\frac{h_0}{h_\theta}\cdot g_0(h_\theta)\Big)] - E[\Big(\frac{\partial}{\partial\theta}\Big|_{\theta=0}\frac{h_0}{h_\theta}\Big)\cdot\Big(g_0(h_\theta)\Big|_{\theta=0}\Big)] \\
&= \frac{\partial}{\partial\theta}\Big|_{\theta=0} E[\frac{h_0}{h_\theta}\cdot g_0(h_\theta)] - E[\Big(-\frac{r(X)}{h_0}\Big)\cdot\Big(g_0(h_0)\Big)] \\
&= \frac{\partial}{\partial\theta}\Big|_{\theta=0} E_X\{E[\frac{T}{h_\theta}\cdot g_0(h_\theta)\Big|X]\} + E[\frac{r(X)}{h_0}\cdot g_0(h_0)] \\
&= \frac{\partial}{\partial\theta}\Big|_{\theta=0} E[\frac{T}{h_\theta}\cdot E[Y\mid T=1,h_\theta]] + E[\frac{r(X)}{h_0}\cdot g_0(h_0)] \\
&= \frac{\partial}{\partial\theta}\Big|_{\theta=0} E[\frac{TY}{h_\theta}] + E[\frac{r(X)}{h_0}\cdot g_0(h_0)] \\
&= \frac{\partial}{\partial\theta}\Big|_{\theta=0} E[\frac{h_0 E[Y\mid T=1,X]}{h_\theta}] + E[\frac{r(X)}{h_0}\cdot g_0(h_0)] \\
&= E[\Big(\frac{\partial}{\partial\theta}\Big|_{\theta=0}\frac{h_0}{h_\theta}\Big)E[Y\mid T=1,X]] + E[\frac{r(X)}{h_0}\cdot g_0(h_0)] \\
&= E[\Big(-\frac{r(X)}{h_0}\Big)E[Y\mid T=1,X]] + E[\frac{r(X)}{h_0}\cdot g_0(h_0)] \\
&= E[(-1)\cdot\Big(\frac{E[Y\mid T=1,X]-g_0(h_0)}{h_0(X)}\Big)\cdot r(X)].
\end{aligned}
$$

Because the above conditional expectations were taken over X or h_0, calculation of $d^*_{\Gamma_n}(x)$ which is equal to $E[\partial g_0(h_0) d^*_{g_n}(h_0) R(X)^T]$ follows the same derivation, and we finally have

$$
\begin{aligned}
& d^*_{\Gamma_n}(x) = E[\partial g_0(h_0) d^*_{g_n}(h_0) R(X)^T] Q^{-1}_{L(n)} R(x) \\
&= (-1)E\Big\{\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)} E[B(h_0(X))^T] Q_K^{-1} B[h_0(X_i)] R(X)^T\Big\} Q_L^{-1} R(x).
\end{aligned}
$$

In summary, we have

$$\triangle_{\varphi}(O_i, h_0)[d^*_{\Gamma_n}]$$
$$= (-1)E\Big\{\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}E[B(h_0(X))^T]Q_K^{-1}B[h_0(X_i)]R(X)^T\Big\}Q_L^{-1}R(X_i)[T_i-h_0(X_i)]$$

and

$$\triangle_{\psi}(O_i, g_0, h_0)[d^*_{g_n}(h_0)]$$
$$= E[B(h_0(X))^T]Q_K^{-1}B[h_0(X_i)]\frac{T_i}{h_0(X_i)}\{Y_i - g_0[h_0(X_i)]\}.$$

By Section 3 of [70], $E[B(h_0(X))^T]Q_K^{-1}B[h_0(X_i)]$ is the population projection of a constant 1 onto the space $\mathcal{G}_n$ spanned by $B[h_0] = B_K[h_0] = [p_1(h_0), \ldots, p_K(h_0)]^T$. As K gets big, under Assumption 1.2(ii) that bias from the series approximation is asymptotically negligible, we have that $E[B(h_0(X))^T]Q_K^{-1}B[h_0(X_i)]$ in the limit approaches the respective projections on the entire sequence $[p_1(h_0), p_2(h_0), \ldots]^T$, which is the conditional expectation $E[1|h_0] = 1$.

Similarly, $E\Big\{\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}E[B(h_0(X))^T]Q_K^{-1}B[h_0(X_i)]R(X)^T\Big\}Q_L^{-1}R(X_i)$ is the population projection of $\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}E[B(h_0(X))^T]Q_K^{-1}B[h_0(X_i)]$ onto the space $\mathcal{H}_n$ spanned by $R[x] = R_L[x] = [r_1(x), \ldots, r_L(x]^T$. As both L and K get big, under Assumption 1.1(iii) and 1.2(ii) that bias from the series approximation is asymptotically negligible, we have that

$$\lim_{L\to\infty}\lim_{K\to\infty} E\Big\{\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}\Big(E[B(h_0(X))^T]Q_K^{-1}B[h_0(X_i)]\Big)R(X)^T\Big\}Q_L^{-1}R(X_i)$$
$$= \lim_{L\to\infty} E\Big\{\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}\Big(E[1 \mid h_0]\Big)R(X)^T\Big\}Q_L^{-1}R(X_i)$$
$$= \lim_{L\to\infty} E\Big\{\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}R(X)^T\Big\}Q_L^{-1}R(X_i)$$
$$= E\Big\{\frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}\Big|X\Big\}$$
$$= \frac{E[Y|T=1,X]-g_0[h_0(X)]}{h_0(X)}$$

Therefore, as $n \to \infty$, both K and L goes to infinity, and we have

$$\triangle_{\varphi}(O_i, h_0)[d^*_{\Gamma_n}] = (-1)\frac{E[Y|T=1, X_i] - g_0[h_0(X_i)]}{h_0(X_i)}[T_i - h_0(X_i)] + o(n^{-1/2})$$
$$= (-1)\frac{E[Y|T=1, X_i] - E[Y|T=1, h_0(X_i)]}{h_0(X_i)}[T_i - E[Y|T=1, h_0(X_i)]] + o(n^{-1/2})$$

and

$$\triangle_{\psi}(O_i, g_0, h_0)[d^*_{g_n}] = \frac{T_i}{h_0(X_i)}\{Y_i - g_0[h_0(X_i)]\} + o(n^{-1/2})$$
$$= \frac{T_i}{h_0(X_i)}\{Y_i - E[Y|T=1, h_0(X_i)]\} + o(n^{-1/2})$$

□

Proof of Corrolary 3.2

Corrolary 3.2 Under Assumptions 1.1-1.4, $\hat{E}[Y(1)]$ is asymptotically linear with

$$\hat{E}[Y(1)] - E[Y(1)] = \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{T_i}{h(X_i)}(Y - E[Y|T=1, X_i]) + E[Y|T=1, X_i] - E[Y(1)]\Big\} + \|v^*_n\|_{sd} o(n^{-1/2}).$$

Consequently, we have

$$\sqrt{n}\Big(\hat{E}[Y(1)] - E[Y(1)]\Big) \to_d N(0, \sigma^2),$$

where $\sigma^2 = \mathrm{Var}\{\frac{T}{h(X)}(Y - E[Y|T=1, X]) + E[Y|T=1, X] - E[Y(1)]\}$.

Proof. For simplicity, we introduce the following notation

$$\rho_n[\hat{g}_n(\hat{h}_n)] = P_n[\hat{g}_n(\hat{h}_n)] = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(\hat{h}_n(X_i))$$
$$\rho[g_0(h_0)] = P_0[g_0(h_0)] = \int g_0(h_0(X))dP(x),$$

then we have

$$\hat{E}[Y(1)] - E[Y(1)] = P_n[\hat{g}_n(\hat{h}_n)] - P_0[g_0(h_0)]$$
$$= P_n\{g_0(h_0) - P_0[g_0(h_0)]\} + P_0[\hat{g}_n(\hat{h}_n) - g_0(h_0)] + (P_n - P_0)[\hat{g}_n(\hat{h}_n) - g_0(h_0)]$$
$$= P_n\{g_0(h_0) - P_0[g_0(h_0)]\}$$
$$+ P_n\Big\{\triangle_\psi(O, g_0, h_0)[d^*_{g_n}] + \triangle_\varphi(O, h_0)[d^*_{\Gamma_n}]\Big\} + \|v^*_n\|_{sd} \cdot o(n^{-1/2}) \quad \text{by Thm 3.2}$$
$$+ (P_n - P_0)[\hat{g}_n(\hat{h}_n) - g_0(h_0)]$$
$$= P_n\Big\{g_0(h_0) - \rho[g_0(h_0)]$$
$$+ \frac{T}{h_0(X)}\{Y - g_0[h_0(X)]\} + (-1)\frac{E[Y|T=1,X] - g_0[h_0(X)]}{h_0(X)}[T_i - h_0(X)]\Big\} + \|v^*_n\|_{sd} \cdot o(n^{-1/2})$$

with $(P_n - P_0)[\hat{g}_n(\hat{h}_n) - g_0(h_0)] = o(n^{-1/2})$ by similar empirical process argument in Lemma B.4

$$= \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{T_i}{h(X_i)}(Y - E[Y|T=1,X_i]) + E[Y|T=1,X_i] - E[Y(1)]\Big\} + \|v^*_n\|_{sd} o(n^{-1/2})$$

By the central limit theorem,

$$\sqrt{n}\Big(\hat{E}[Y(1)] - E[Y(1)]\Big) \to_d N(0, \sigma^2),$$

where $\sigma^2 = \mathrm{Var}\{\frac{T}{h(X)}(Y - E[Y|T=1,X]) + E[Y|T=1,X] - E[Y(1)]\}$.

□