EMPOWERING RCT WITH MULTI-SITE MULTI-SOURCE RWD: A STATISTICAL LEARNING PERSPECTIVE

Dazheng Zhang

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2025

Supervisor of Dissertation

Yong Chen, Professor of Biostatistics

Graduate Group Chairperson

Russell Shinohara, Professor of Biostatistics

Dissertation Committee

David A. Asch, Professor of Medicine, Professor of Medical Ethics and Health Policy, Professor of Health Care Management, Professor of Operations, Information and Decisions
Douglas E. Schaubel, Professor of Biostatistics
David Wolk, Professor of Neurology
Wei Yang, Professor of Biostatistics

EMPOWERING RCT WITH MULTI-SITE MULTI-SOURCE RWD: A STATISTICAL LEARNING PERSPECTIVE

ACKNOWLEDGEMENT

I am grateful to Dr. Yong Chen for his supervision, mentorship, and support throughout my academic journey. From the very first day, Dr. Chen has believed in my potential, consistently providing insightful guidance and encouraging me to explore new ideas and methodologies. His dedication to excellence and enthusiasm for research have not only inspired my passion for science but also fostered my personal growth and confidence as an independent researcher. The lessons I've learned from his leadership, patience, and thoughtful mentorship will remain invaluable throughout my career.

I wish to express my sincere gratitude to my thesis committee members, whose guidance have significantly contributed to my academic growth.

I would also like to extend my heartfelt appreciation to the members of the Center for Health AI and Synthesis of Evidence (CHASE). Thank you all for creating such a joyful, supportive, and intellectually stimulating environment. The countless discussions, exchanges of ideas, and shared enthusiasm for research have enriched my learning experience and made my time in the lab truly enjoyable.

Finally, my deepest gratitude goes to my family and friends for their constant love, support, and encouragement. Their belief in me and understanding throughout this journey have provided the strength and motivation I needed to succeed. Thank you all for standing by my side, always.

ABSTRACT

EMPOWERING RCT WITH MULTI-SITE MULTI-SOURCE RWD: A STATISTICAL LEARNING PERSPECTIVE

Dazheng Zhang

Yong Chen

Randomized controlled trials (RCTs) are recognized for their ability to minimize confounding through randomization, thereby establishing robust causal inferences in clinical research. However, the stringent protocols and limited sample sizes typical of RCTs can restrict statistical efficiency and the scope of inference. In contrast, real-world data (RWD), such as electronic health records (EHRs), capture broader and more representative populations but introduce challenges including bias, missing data, and the lack of randomization. Integrating these data sources effectively can enhance both statistical efficiency and external validity, yet it requires overcoming issues such as heterogeneity across clinical sites, bias from unmeasured confounding, and model shifts between trial and real-world populations.

The overarching goal of this dissertation is to develop and validate robust analytical frameworks that address the limitations of using RWD in federated learning and its integration with RCTs. This work is organized around three interconnected aims. The first aim focuses on the development of one-shot distributed algorithms designed to analyze competing risks data across decentralized distributed research networks, explicitly addressing data heterogeneity to generate high-quality real-world evidence that informs clinical decision-making. In settings where event rates are extremely rare, this aim further introduces a distributed Firth-corrected inference method utilizing a two-round communication strategy to improve estimation accuracy.

The second aim addresses the challenge of time-varying effect of unmeasured confounding by proposing a novel negative control-calibrated difference-in-differences (NC-DiD) methodology. By integrating negative control outcomes, this approach systematically detects, quantifies, and corrects for

biases, thereby enhancing the validity of causal inferences drawn from EHR data. Its application to evaluate racial and ethnic differences in post-acute sequelae of COVID-19 among pediatric populations demonstrates significant improvements in bias correction and estimation reliability compared to conventional methods.

The third aim combines RCTs with RWD to enhance statistical efficiency and accelerate clinical trials. This is achieved through the development of the Negative Control-Calibrated Digital Twin framework, which constructs individualized digital twins from RWD—calibrated via negative control outcomes—to mitigate model shift bias between trial and real-world populations. Empirical validation using neuroimaging data from the SPRINT-MIND trial alongside data from the iSTAGING Consortium confirms significant enhancements in statistical efficiency and robustness.

Overall, the contributions of this dissertation provide novel methodological insights and practical tools that advance the integration of RCT and RWD, ultimately improving causal inference and the external validity of clinical research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1. Motivation for Chapter 2

Federated research networks provide the potential for collaborative analysis of large-scale, decentralized data without sharing individual-level information. However, the presence of heterogeneity across clinical sites can significantly bias findings. Chapter 2 addresses the urgent need for computationally efficient, privacy-preserving methods capable of analyzing competing risks data across decentralized clinical sites. By developing a novel "one-shot distributed algorithm" (ODACoRH), this chapter aims to improve the accuracy and efficiency of federated analyses. The practical relevance and utility of this method are demonstrated through its application to pediatric Crohn's disease and otitis media, establishing its validity and broad applicability.

1.2. Motivation for Chapter 3

Although federated learning methods have advanced, biases remain problematic, particularly when dealing with rare-event data. Rare diseases, despite low prevalence, demand accurate and unbiased estimations due to their clinical severity and implications for patient care. Chapter 3 introduces a bias-corrected federated learning framework employing the distributed Firth estimator, specifically tailored to rare-event contexts. This methodology significantly mitigates first-order bias in distributed inference, enhancing the reliability of federated learning results in rare disease scenarios. The practical relevance is demonstrated through a critical application to eosinophilic granulomatosis with polyangiitis, showcasing improved accuracy and reduced bias compared to existing federated learning methods.

1.3. Motivation for Chapter 4

Unmeasured confounding remains a critical obstacle when drawing causal inferences from observational RWD. Traditional difference-in-differences (DiD) analyses hinge upon the assumption of parallel trends, which is frequently violated in practice, potentially compromising validity. Chapter 4 addresses this by proposing a novel Negative Control-Calibrated Difference-in-Differences (NC-

DiD) methodology. This approach uses negative control outcomes to systematically detect, quantify, and correct for unmeasured confounding biases, significantly improving causal inference validity. Its application to assessing racial and ethnic differences in COVID-19 sequelae among pediatric populations demonstrates substantial methodological improvement and its significant potential to inform clinical and public health interventions.

1.4. Motivation for Chapter 5

Post-acute sequelae of SARS-CoV-2 infection (PASC), or "long COVID," significantly impacts pediatric populations. Yet, limited knowledge exists regarding racial and ethnic differences in PASC among children and adolescents. This chapter addresses this knowledge gap using data from the RECOVER Initiative, encompassing diverse pediatric cohorts across multiple institutions. Employing a novel Negative Control-Calibrated Difference-in-Differences (NC-DiD) method, the research identifies racial/ethnic differences in pediatric PASC attributable specifically to COVID-19.

1.5. Motivation for Chapter 6

RCTs, despite their robustness, often struggle with limited sample sizes and external validity. Integrating RWD into RCTs can theoretically alleviate these limitations but introduces risks of model shift and related biases. Chapter 5 innovatively proposes a Negative Control-Calibrated Digital Twin framework, designed to rigorously combine RCTs with multi-site, multi-source RWD. By constructing individualized "digital twins" calibrated through negative control outcomes, this method significantly enhances statistical efficiency and robustness against model-shift bias. Validation using neuroimaging data from the SPRINT-MIND trial and the iSTAGING Consortium demonstrates remarkable improvements in efficiency and external validity, indicating strong potential for future clinical trial designs.

Collectively, this dissertation not only presents novel statistical methodologies but also delivers practical tools and insights for the effective integration of RCTs and RWD, ultimately enhancing causal inference, efficiency, and generalizability in clinical research.

CHAPTER 2

ONE-SHOT DISTRIBUTED ALGORITHMS FOR ADDRESSING HETEROGENEITY IN COMPETING RISKS DATA ACROSS CLINICAL SITES

2.1. Introduction

Real-world evidence (RWE) refers to data collected in environments outside of the controlled setting of randomized clinical trials, including electronic health records (EHRs), claims databases, and patient registries (Shea and Hripcsak, 2010; Sherman et al., 2016; Friedman et al., 2010). For example, EHRs comprise digital records of patient clinical information that clinicians can use to make informed decisions, generate knowledge, and predict future health risks (Sherman et al., 2016). In recent years, there has been a surge in the development of RWE distributed research networks (DRN), such as the Observational Health Data Sciences and Informatics (OHDSI) (Hripcsak et al., 2015) and the Stakeholders, Technology, and Research Clinical Research Network (STAR-CRN) (STAR Clinical Research Network, n.d.). These networks provide richer data sources and larger sample sizes, allowing for better generalization and higher statistical power compared with using the data from a single hospital. Additionally, the temporal information (i.e., time-to-event data) of clinical outcomes or events presented in these networks enables the identification of disease progression patterns over time, providing clinicians with valuable insights for timely personalized medication, and assessing the timing and probability of adverse events in drug evaluation.

RWE offers a valuable opportunity to perform rigorous time-to-event analyses, serving as a crucial resource to inform and guide clinicians in their practice. However, using data obtained from multiple clinical sites to study time-to-event outcomes can pose several statistical challenges. Firstly, individual-level data cannot be shared across decentralized networks (e.g., OHDSI), and only aggregated data are allowed to be shared between sites or hospitals. Secondly, transferring and retrieving data across sites requires human intervention, which can be time-consuming and inefficient. To address the two challenges above, a number of communication-efficient federated algorithms have been proposed (Shu et al., 2020; Duan et al., 2018, 2020a,b; Edmondson et al., 2021, 2022; Liu et al.,

2022; Tong et al., 2022; Luo et al., 2022b; Wu et al., 2023; Tong et al., 2024; Zhang et al., 2024). However, the third challenge which has not been addressed in the literature until recently is data heterogeneity. In particular, patient characteristics such as age or ethnicity can vary based on factors such as geographical location, cultural norms, and socioeconomic status, leading to biased results if ignored (Tong et al., 2019; Duan et al., 2022). To overcome these challenges in data sharing, communication cost, and heterogeneity, a new algorithm for time-to-event analysis called ODACH has been developed (Luo et al., 2022a). It is a one-shot distributed algorithm for the Cox model that accounts for between-site heterogeneous baseline hazard function without sharing patient-level data.

Current algorithms used to study time-to-event outcomes in medicine have been focusing on a single event. On the other hand, EHRs capture multiple types of events, where studying these events concurrently offers a comprehensive understanding of patient care and informs better clinical decisions. Ignoring multiple events and fitting a standard univariate time-to-event model can introduce bias (Haneuse and Lee, 2016), making it important to study multiple events simultaneously. Pediatric Crohn's Disease (PCD) provides an exceptional example of why multiple events should be considered, as it has three different phenotypes: inflammatory, stricturing, and penetrating (Levine et al., 2011). Patients with PCD often progress from the inflammatory phenotype to stricturing or penetrating disease over time. By studying all three phenotypes simultaneously, researchers can better understand the full disease trajectory and the interplay of clinical events.

Rarely, children with PCD may experience both stricturing and penetrating phenotypes concurrently. The occurrence of the first stricturing or penetrating phenotype is often mutually exclusive, and the two phenotypes can preclude each other. Therefore, it is necessary to use competing risk models (Fine and Gray, 1999; Zhou et al., 2012) to account for the feature of multiple events in the analysis. Ignoring the competing risks of multiple events can lead to biased estimates of the cumulative incidence function for the primary event in survival analysis. (Austin et al., 2016) This is because the occurrence of one event can affect the probability of experiencing another event. For instance, if a patient experiences a penetrating phenotype, they may be less likely to experience a

stricturing phenotype.

To the best of our knowledge, no one-shot algorithms have been developed for multiple events, especially for the competing risk data in heterogeneous multi-site settings. Therefore, we developed a heterogeneous baseline hazard function inspired by the competing risks model for clustered data (Zhou et al., 2012). Our proposed algorithm has several advantages as follows: (1) heterogeneity-aware, different sites have their own site-specific subhazard baseline function, (2) capable of handling multi-type events, enables analyzing multiple events simultaneously (3) communication-efficient, only involves one round of communication, (4) aggregated data based, only the aggregated data is transferred across different sites, (5) accurate, the ODACoRH achieves almost identical performance compared to the results if the data were pooled together. Overall, our proposed method provides a new way to handle the competing risks data in a communication-efficient and heterogeneity-aware manner.

To validate the ODACoRH, we apply the algorithm to a cohort of 2,944 patients who experienced PCD between January 2009 and August 2020 in the PEDSnet (Forrest et al., 2014), a national pediatric health learning system involving eleven children's hospitals. A unique strength of our empirical evaluation of the ODACoRH algorithm is that a centralized dataset at PEDSnet is available where decentralized data were pooled together so that the pooled analyses can be conducted as the gold standard for comparison. We have applied the proposed algorithm to an otitis media (OM) study, aiming to further validate the generalizability of the ODACoRH algorithm in handling the between-site heterogeneity. In this OM study, the cohort in obtained within the PEDSnet on 51,365 pediatric patients with OM, who have been given the initial intervention of the tympanostomy tube insertion (TTI).

The extensive simulation studies and the two real-world data cases for the PCD and OM demonstrate that the ODACoRH has an identical performance on par with the pooled analysis in estimating the subdistribution hazard ratio. The proposed algorithm is also highly generalizable, as it can be applied to decentralized datasets and extended to other healthcare settings.

2.2. Methods

We assume a study involving J clinical sites and consider the relationship between risk factors Z and the observed time-to-event outcome X, which may be censored at time C. Figure 2.1 illustrates the data structure for our proposed algorithm ODACoRH.

Under the competing risks setting, suppose that there are K types of events denoted by Π, where $\Pi = k$ $(k = 1, \ldots, K)$ indicates that the type k event happens at time T_k. We denote T as the time of a composite time point encompassing multiple event types, represented as the minimum value among T_1 to T_K. The minimum value of T and C is taken as the observed time X, while Δ is a censoring indicator, with $\Delta = 1$ indicating an event and $\Delta = 0$ indicating censoring.

We observe n_j patients for the jth site and the total number of patients across J sites is $n = \sum_{j=1}^{J} n_j$. The observed data for the ith patient from the jth site are denoted by $\{X_{ij}, \Delta_{ij}, \Pi_{ij}, Z_{ij}\}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J.$

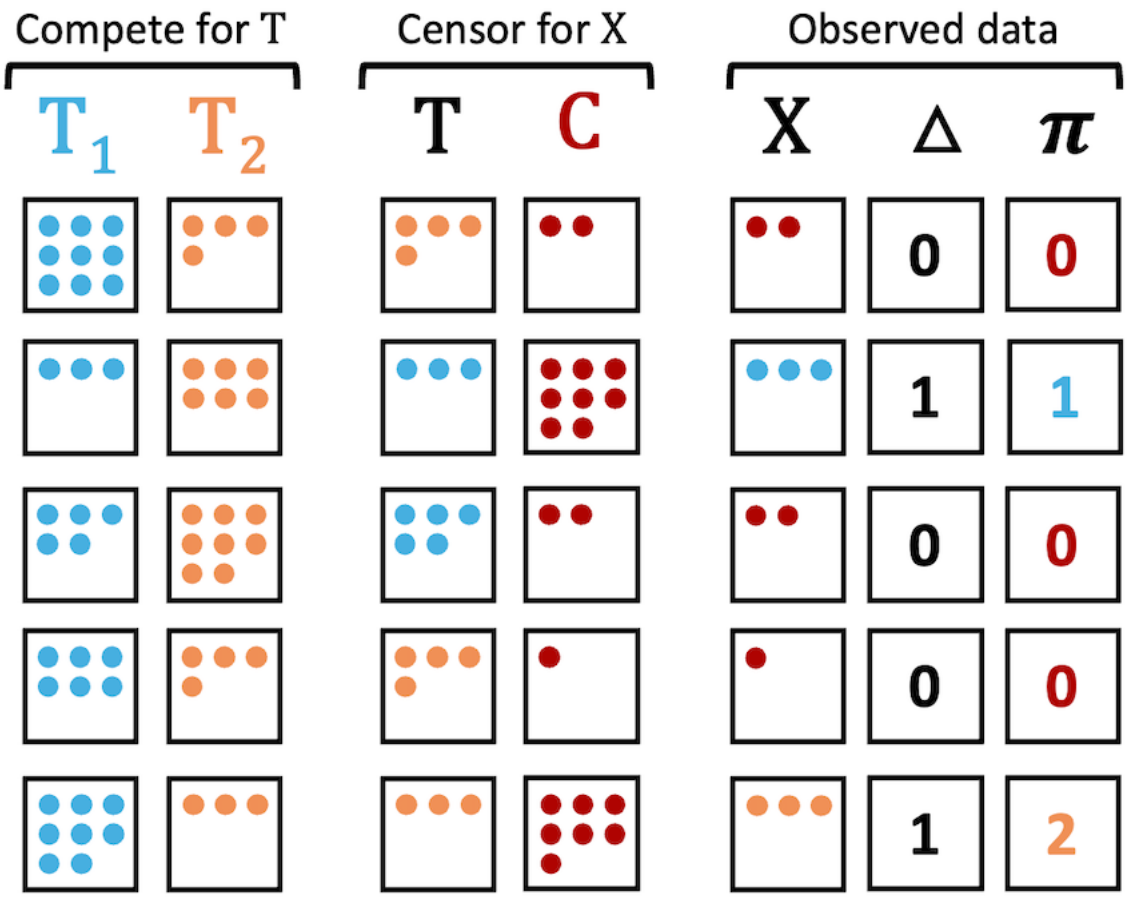

Figure 2.1: Data structure and notation for competing risks data. T_1 and T_2 denote the time of event 1 and event 2 and compete for event time T. T is potentially censored by C and formulates the observed time X. Δ is an indicator for censoring. Type Π is an indicator to denote which event happens. The data structure for the data we observed is (X, Δ, Π).

Subdistribution Hazard Model. We assume that the model has the subdistribution hazard $\lambda_{ijk}(t)$ for the ith patient from the jth site experiencing the kth type of event, which is proportional to the between-site heterogeneous baseline subdistribution hazard $\lambda_{0jk}(t)$ for the type k event. Specifically, we have the equation:

$$\lambda_{ijk}(t) = \lambda_{0jk}(t) \exp\left(Z_{ij}^T \beta_k\right). \tag{2.1}$$

Here, we assume that the effect size of the kth type event, represented by β_k, is constant across sites. The vector β_k can be interpreted as the logarithm of the subhazard ratio.

Gold Standard Method from Pooling Data. We can obtain the pooled estimate $\hat{\beta}_k$ by maximizing the likelihood L_k of Equation 2.1 for the type k event, if the data from J sites can be pooled together. The exact form of L_k is provided in the Supplementary Materials. However, due to the inability to share individual-level data from different sites, it is challenging to obtain the pooled estimate. This estimate is referred to as the gold standard estimate, as it enjoys the best accuracy and most efficient properties.

Meta-Analysis Method. When the data cannot be pooled together, a common solution is the meta-analysis method (Borenstein et al., 2021; Lake et al., 2024; Liu et al., 2025; Tang et al., 2025). The estimate obtained from the meta-analysis method, denoted by $\bar{\beta}_k$, is shown as

$$\bar{\beta}_k = \Big(\sum_{j=1}^{J} V_{jk}^{-1}\Big)^{-1} \sum_{j=1}^{J} \Big(V_{jk}^{-1}\, \hat{\beta}_{jk}\Big), \tag{2.2}$$

where $\hat{\beta}_{jk}$ is found by maximizing the local likelihood function for the jth site's type k event and V_{jk} is the variance of $\hat{\beta}_{jk}$. The variance of the meta-analysis can be calculated from Equation (A.11) in the Supplementary Materials.

Proposed ODACoRH Algorithm. Our proposed algorithm, ODACoRH, is designed to require only one round of communication. We assume, without loss of generality, that the first site is the lead site ($j = 1$). Inspired by the communication-efficient surrogate likelihood (Jordan et al., 2019), we utilize a surrogate likelihood to approximate the likelihood function of Equation (2.1). The surrogate likelihood function of the kth type of events is given by

$$\begin{aligned}\tilde{L}_k(\beta_k) = L_{1k}(\beta_k) &+ \big(\nabla L_k(\bar{\beta}_k) - \nabla L_{1k}(\bar{\beta}_k)\big)^T(\beta_k - \bar{\beta}_k) \\ &+ \tfrac{1}{2}\,(\beta_k - \bar{\beta}_k)^T\big(\nabla^2 L_k(\bar{\beta}_k) - \nabla^2 L_{1k}(\bar{\beta}_k)\big)(\beta_k - \bar{\beta}_k),\end{aligned} \tag{2.3}$$

where $L_{1k}(\beta_k)$ is the local likelihood function of the lead site, and ∇^e denotes the eth-order gradient of the likelihood functions. The term $\bar{\beta}_k$ is the meta-analysis estimator shown in Equation (2.2). During the first round of communication, the ODACoRH algorithm computes $\bar{\beta}_k$. In the second round of communication, the ODACoRH algorithm calculates the estimate $\tilde{\beta}_k$ by maximizing Equa-

tion (2.3).

As shown in Equation (2.3), the construction of $\widetilde{L}_k(\beta_k)$ requires the individual-level data from the lead site and the aggregated data (first-order gradient of the pooled likelihood function, $\nabla L_k(\bar{\beta}_k)$, and second-order gradient, $\nabla^2 L_k(\bar{\beta}_k)$, from the pooled likelihood function). The exact equations of $L_{1k}(\beta_k)$, $\nabla L_{1k}(\bar{\beta}_k)$, $\nabla^2 L_{1k}(\bar{\beta}_k)$, $\nabla L_k(\bar{\beta}_k)$, and $\nabla^2 L_k(\bar{\beta}_k)$ are shown in the Supplementary Materials. From those equations, we have

$$\nabla L_k(\bar{\beta}_k) = \sum_{j=1}^{J} \nabla L_{jk}(\bar{\beta}_k), \tag{2.4}$$

$$\nabla^2 L_k(\bar{\beta}_k) = \sum_{j=1}^{J} \nabla^2 L_{jk}(\bar{\beta}_k). \tag{2.5}$$

This suggests that the first- and second-order gradients of the pooled likelihood function can be viewed as the collection of the first- and second-order gradients of the local likelihood functions from each site. Since the lead site has already calculated the first- and second-order gradients of the likelihood function for itself, we only require the jth partner site ($j = 2, \ldots, J$) to transfer its first- and second-order gradients of the local likelihood function to the lead site. The variance of $\widetilde{\beta}_k$ can be calculated from Equation (A.10) in the Supplementary Materials. We ran the same procedure for all event types from 1 to K. The workflow and the detailed pseudo-algorithm of the ODACoRH are shown in Figure 2.2 and Algorithm 1.

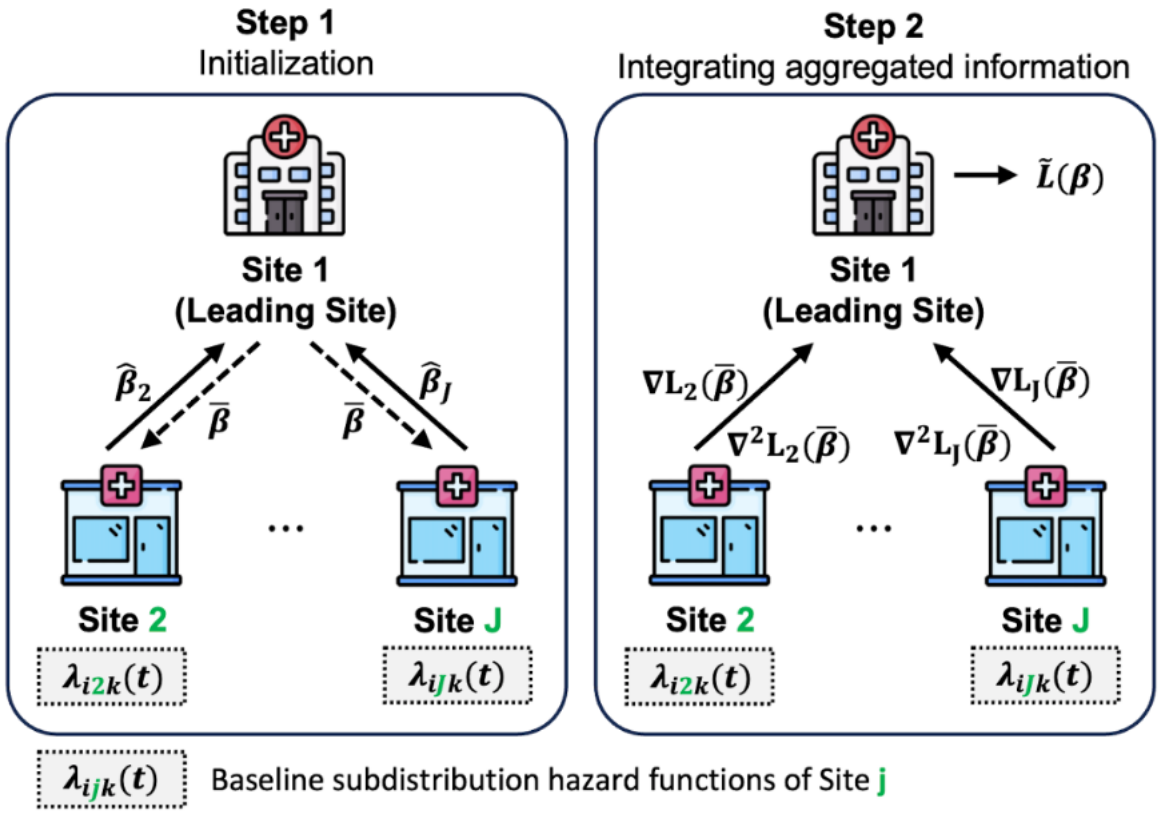

Figure 2.2: The working flow of ODACoRH. The first step is to find the initial value. Each site calculates the local estimate and transfers it to the lead site. The lead site calculates the initial value and sends it back to each site. The second step is to integrate the aggregated information and obtain the surrogate estimate. The lead site collects the summary-level statistics from each site. It assembles the $\widetilde{L}_k(\beta_k)$ and obtains the estimate $\widetilde{\beta}_k$ that maximizes Equation (2.3).

Simulation Settings. To cover a wide spectrum of real-world settings, we conducted extensive numerical experiments to evaluate the performance of ODACoRH. Figure 2.3 displays the heterogeneity of the baseline distribution of type 1 and type 2 events on a discrete time scale. We simulated a population of $n = 10{,}000$ patients distributed across $J = 10$ sites, where two types of events (type 1 and type 2) occurred.

To mimic PCD data, we set the logs of subhazard ratios for type 1 and type 2 events to be 1 for two covariates following uniform distributions, which represent population characteristics such as BMI and age. The values of coefficients β_1 and β_2 are set to be $(1, 1)$. We controlled the event rates of

Algorithm 1 Pseudo-algorithm for the ODACoRH

Initialization:
for site j from 1 to J **do**
    for type k from 1 to K **do**
        Obtain local-site estimate $\hat{\beta}_{jk}$.
    end for
    Transfer $\hat{\beta}_j = \left(\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jK}\right)$ to the lead site.
end for
In the lead site ($j = 1$), calculate the initial estimate $\bar{\beta}$ from the meta-analysis method and broadcast it to the other sites.
Integrate aggregated information:
for site j from 1 to J **do**
    for type k from 1 to K **do**
        Obtain local $\nabla^e L_{jk}\left(\bar{\beta}_k\right)$ for $e = 1, 2$.
    end for
    Transfer $\nabla^e L_j(\bar{\beta}) = \left(\nabla^e L_{j1}(\bar{\beta}_1), \ldots, \nabla^e L_{jK}(\bar{\beta}_K)\right)$ to the lead site.
end for
In the lead site:
for type k from 1 to K **do**
    Assemble $\widetilde{L}_k$ and then obtain the estimate $\widetilde{\beta}_k$ from Equation (2).
end for

type 1 and type 2 to be equal.

To mimic the heterogeneity of the baseline subhazard function of the type 1 event, we generated between-site heterogeneous baseline subhazard functions from the mixture Weibull distribution with the scale parameter ranging from 100 to 200 and a probability mass of 0.5 at infinity, while the shape parameter was fixed at 10. Similarly, the baseline subhazard function of the type 2 event was generated from the Weibull distribution with the scale parameter ranging from 120 to 220, while the shape parameter was fixed at 8.

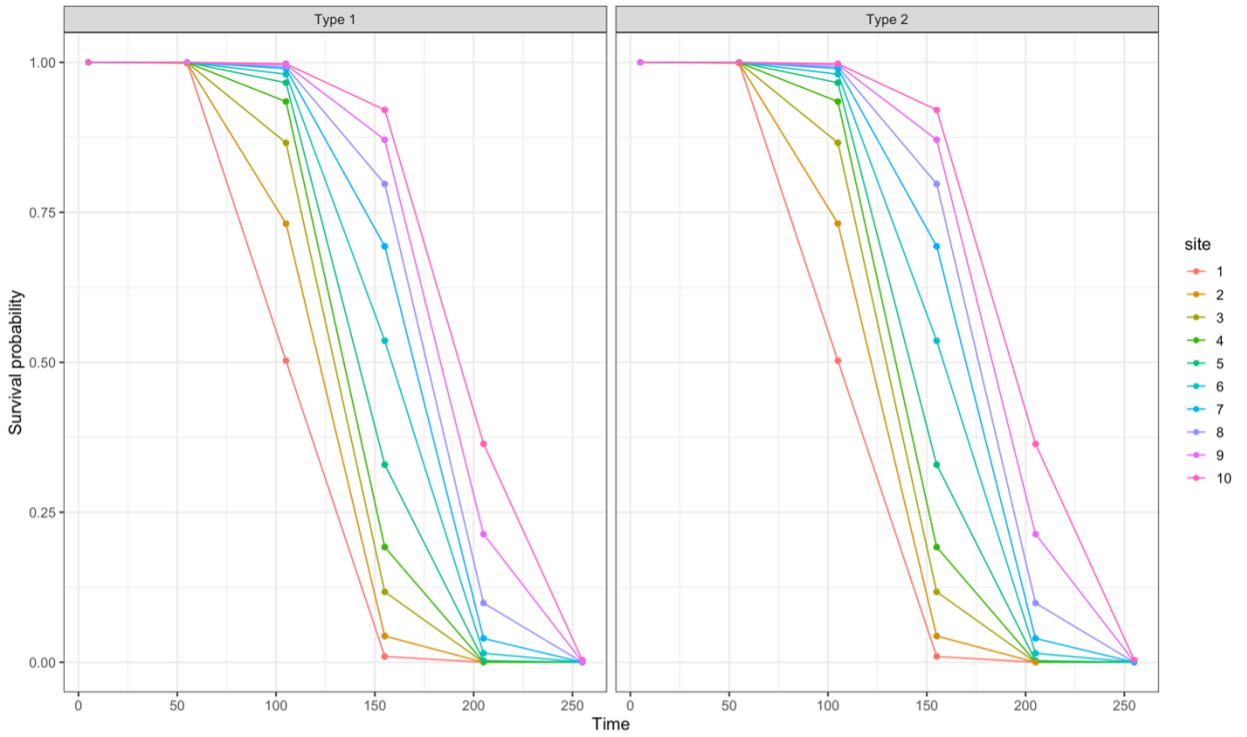

Figure 2.3: Survival probability for underlying heterogeneous subdistribution baseline hazard functions. The left panel shows the type 1 event's baseline subhazard function. The right panel shows the type 2 event's baseline subhazard function. Each color indicates each site. At each time point, the survival probabilities vary much across different sites.

We investigated the performance of ODACoRH for total event rates ranging from the common case (20%) to rare cases (2% and 1%), by generating time of censoring following the appropriate uniform distributions. We compared ODACoRH with the meta-analysis method by calculating the relative bias to the gold standard computed by pooling data together from 250 repetitions of the numerical experiments. For simplicity, we presented the results for the second covariate subhazard ratio for the type 1 event, but the results for other coefficients were similar.

2.3. Results

Simulation results. To cover a wide spectrum of real-world settings, we conducted numerical experiments to evaluate the empirical performance of the ODACoRH algorithm including the settings where the event rates of multiple types of events are relatively rare. To mimic the real-world data, we simulated the time points of two events with two covariates. The details of the simulation

setting were stated in the Methods Section. We compared the ODACoRH with the meta-analysis method for the subhazard ratio for the type-1 event when the type-2 event presents. The pooled estimate utilized the data pooled together from all the sites and thus can be considered as the gold standard.

Figure 2.4 presents the relative bias of results from ODACoRH and the meta-analysis method for the first component of the type 1 event at total event rates of 20%, 2%, and 0.5%. The results from the ODACoRH and meta-analysis about the second component of the type 1 event are shown in the Supplementary Materials. Our ODACoRH algorithm is superior in terms of relative bias, compared to the meta-analysis method. Specifically, under all three scenarios, the relative bias of ODACoRH was close to 0%, indicating that it was equivalent to the gold standard. Conversely, as event rates became rarer, the results from the meta-analysis method exhibited increasingly higher bias. Specifically, at rare event rates, the results from the meta-analysis method had an average relative bias of -11.4%. In contrast, the results from the ODACoRH algorithm had an average relative bias of only -0.2%.

We also assessed the performance of the ODACoRH algorithm under the assumption of site homogeneity, wherein it is assumed that all sites have the same baseline subhazard distribution functions. The results of this evaluation are presented in Supplementary Materials (Figures A.2 and A.3). Under the homogeneous assumption, the meta-analysis exhibits a larger relative bias in comparison to the gold standard (pooled data), as opposed to the ODACoRH algorithm. For example, when the event rate is 0.5% for both type 1 and type 2 events, the meta-analysis of the coefficient for the first covariate demonstrates a mean relative bias of -7.5%, whereas the ODACoRH yields a mean relative bias of -0.7%.

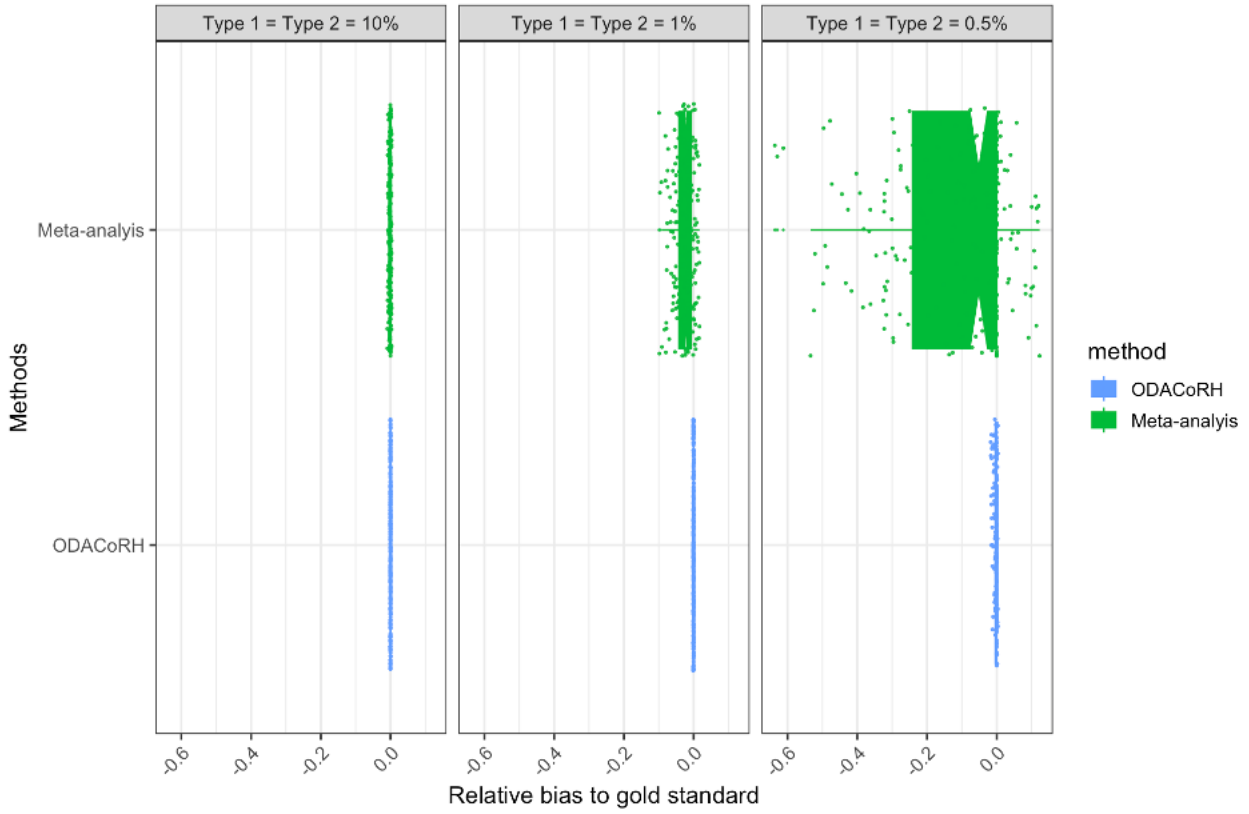

Figure 2.4: The relative bias for ODACoRH vs meta-analysis algorithms to the pooled estimate. In reality, the data cannot be pooled together and thus the pooled estimate is considered as the gold standard. The green boxplot shows the meta-analysis method. The blue boxplot is our proposed algorithm ODACoRH.

Evaluation of Impacts of Risk Factors Associated with the Penetrating or Stricturing Behaviors Among Children with PCD.

To evaluate the performance of our algorithms on real-world data, we conducted a study on PCD using centralized data extracted from PEDSnet. Our study included a total of 2,944 patients from four clinical sites in PEDSnet who experienced two mutually exclusive behaviors: stricturing and penetrating behaviors. To be included in the study, patients needed to have been observed at least once at a PEDSnet institution between January 2009 and August 2020 and be under 19 years old at the start date of their pediatric Crohn's disease. The event rate for stricturing behavior in this cohort was 9.4%, and the event rate for penetrating behavior was 11.3%. The demographic characteristics of the patients are summarized in Table 2.1.

Table 2.1: Characteristics of patients for PCD stratified by sites.

	A (N=1128)	B (N=343)	C (N=613)	D (N=860)	Overall (N=2944)
BMI					
normal	767 (68.0%)	196 (57.1%)	387 (63.1%)	554 (64.4%)	1904 (64.7%)
mild malnutrition	211 (18.7%)	79 (23.0%)	119 (19.4%)	156 (18.1%)	565 (19.2%)
moderate malnutrition	88 (7.8%)	31 (9.0%)	70 (11.4%)	98 (11.4%)	287 (9.7%)
severe malnutrition	62 (5.5%)	37 (10.8%)	37 (6.0%)	52 (6.0%)	188 (6.4%)
ESR (mm/h)					
Mean (SD)	28.6 (23.4)	27.7 (25.1)	33.5 (25.3)	28.1 (24.2)	29.4 (24.3)
Median [Min, Max]	24.0 [0, 134]	18.0 [1.00, 115]	28.0 [1.00, 130]	21.0 [1.00, 133]	24.0 [0, 134]
Hemoglobin (g/dl)					
Mean (SD)	11.4 (1.81)	12.3 (2.06)	11.5 (1.87)	11.6 (1.74)	11.6 (1.86)
Median [Min, Max]	11.6 [4.10, 17.6]	12.4 [4.60, 17.8]	11.4 [3.80, 17.8]	11.6 [4.60, 17.3]	11.6 [3.80, 17.8]
Albumin (g/dl)					
Mean (SD)	3.85 (0.603)	3.88 (0.579)	3.79 (0.590)	3.78 (0.647)	3.82 (0.612)
Median [Min, Max]	3.90 [1.30, 5.40]	4.00 [2.10, 5.00]	3.80 [1.60, 5.40]	3.80 [1.60, 5.20]	3.90 [1.30, 5.40]
Age (Years)					
Mean (SD)	13.0 (3.7)	13.0 (3.5)	13.6 (3.5)	13.4 (3.2)	13.2 (3.5)
Median [Min, Max]	13.6 [2.0, 18.9]	13.5 [2.2, 18.8]	14.2 [2.5, 18.9]	13.9 [2.3, 18.9]	13.8 [2.0, 18.9]
Sex					
Female	472 (41.8%)	141 (41.1%)	263 (42.9%)	388 (45.1%)	1264 (42.9%)
Male	656 (58.2%)	202 (58.9%)	350 (57.1%)	472 (54.9%)	1680 (57.1%)
Race					
White	817 (72.4%)	241 (70.3%)	491 (80.1%)	580 (67.4%)	2129 (72.3%)
Black	129 (11.4%)	18 (5.2%)	67 (10.9%)	105 (12.2%)	319 (10.8%)
Hispanic and others/unknown	182 (16.1%)	84 (24.5%)	55 (8.9%)	175 (20.3%)	496 (16.9%)

We fitted the ODACoRH algorithm, as well as the meta-analysis method, and compared their results to the pooled estimate. We first considered stricturing behavior as the primary outcome and penetrating behavior as the competing risk outcome, and vice versa. Figure 2.5 displays the effect sizes (in the log scale of subdistribution hazard ratio) for the risk factors associated with stricturing behavior and their confidence intervals.

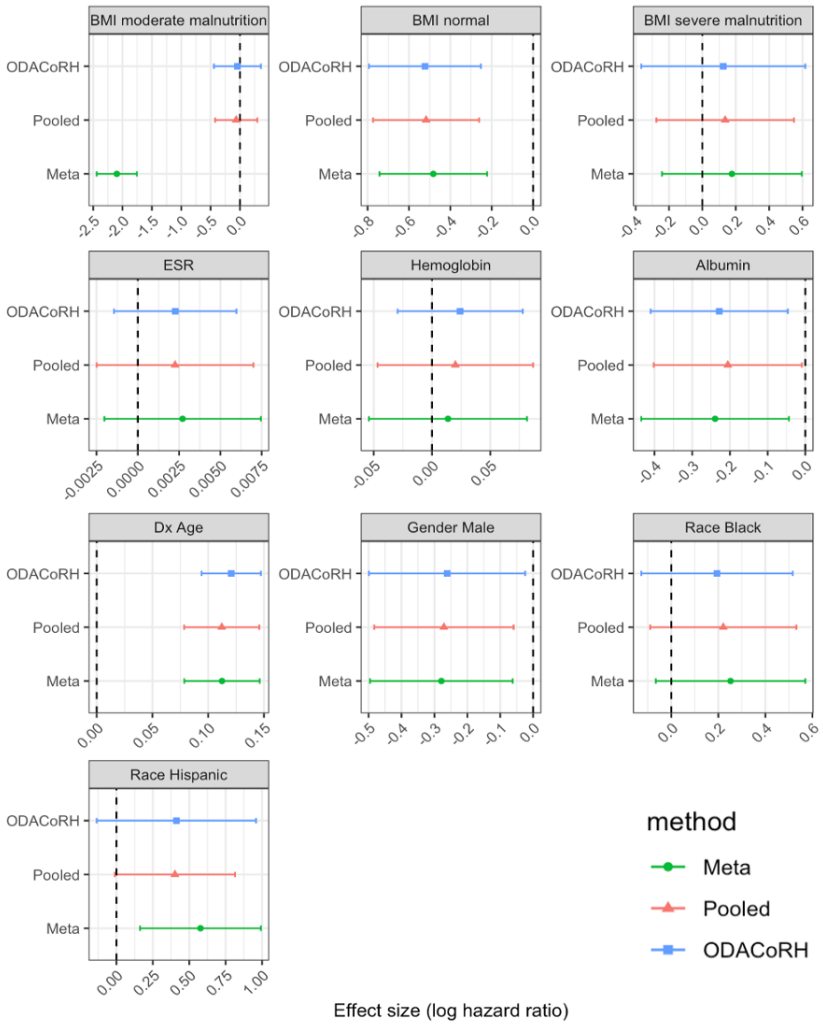

Figure 2.5: The effect size of risk factors associated with stricturing behavior. The effect size is shown in the log scale of the subhazard ratio. The red line indicates the pooled estimate from pooling data together. The green line indicates the meta-analysis method. The blue line indicates the ODACoRH. Each facet shows each potential risk factor. BMI is cut into four categories: normal, mild malnutrition (baseline category), moderate malnutrition, and severe malnutrition. ESR refers to the erythrocyte sedimentation rate. Dx Age means the age at the diagnosis. Race is separated into the following categories: black, Hispanic, and white (baseline category).

Both the pooled estimate and ODACoRH identified four risk factors for stricturing behavior, including normal BMI, lower albumin at diagnosis, older age at diagnosis, and male sex. The ODACoRH yielded results that were identical to the pooled estimate. In contrast, the meta-analysis method produced a large relative bias, leading to misleading interpretations. For instance, the meta-analysis

method suggested that BMI moderate malnutrition was statistically significant ($p<0.001$), while both ODACoRH ($p=0.83$) and pooled estimate ($p=0.73$) found it to be not statistically significant. Figure 2.6 displays the results of ODACoRH for penetrating behavior, where we identified the risk factors to be higher ESR, lower hemoglobin, older age at diagnosis, and male sex. The ODACoRH yielded results that were almost identical to the pooled estimate.

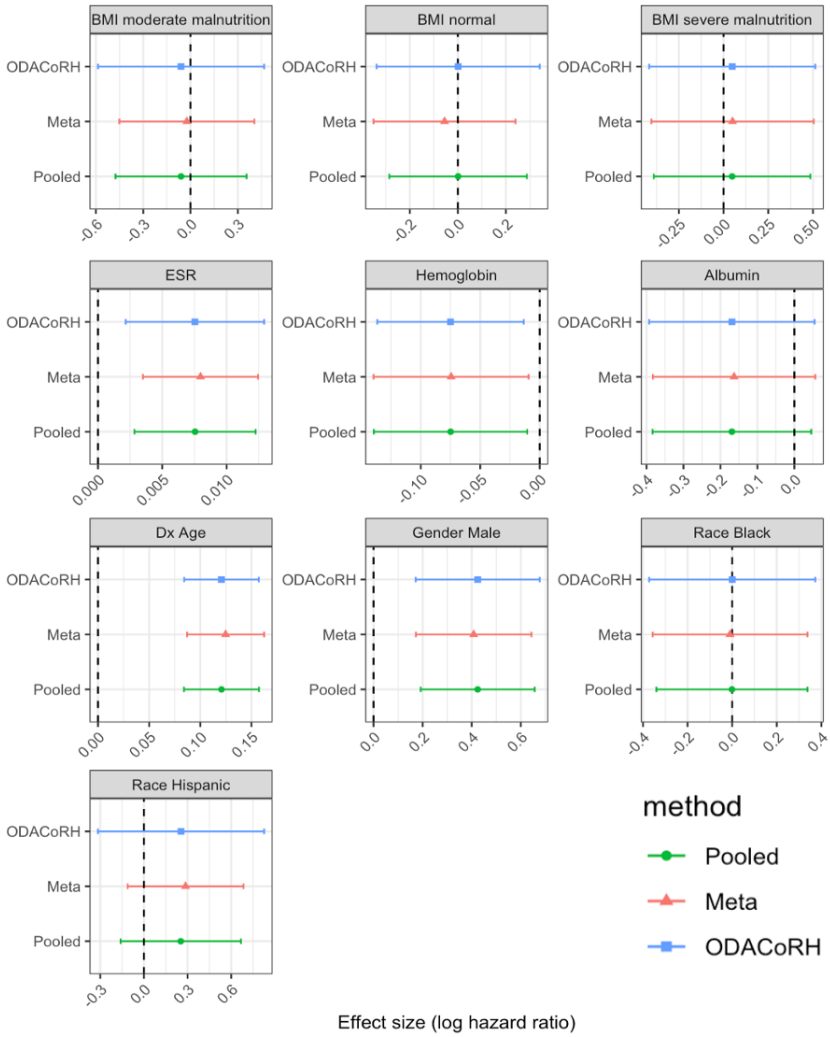

Figure 2.6: Effect size of risk factors associated with penetrating behavior. The effect size is shown in the log scale of the subhazard ratio. The red line indicates the pooled method. The green line indicates the meta-analysis method. The blue line indicates the ODACoRH. Each facet shows each potential risk factor. BMI is cut into four categories: normal, mild malnutrition (baseline category), moderate malnutrition, and severe malnutrition. ESR refers to the erythrocyte sedimentation rate. Dx Age means the age at the diagnosis. Race is separated into the following categories: black, Hispanic, and white (baseline category).

To provide further clinical insights into the risk factors identified by the ODACoRH algorithm, we present a comparison of effect sizes from the ODACoRH algorithm and the RISK study in Table 2.2 (Kugathasan et al., 2017). The RISK study prospectively investigated newly diagnosed PCD patients with an inflammatory phenotype at diagnosis to identify risk factors associated with complicated disease behavior. As shown in Table 2.2, we found that increased age was a risk factor for the development of penetrating phenotype, which is consistent with the findings of the RISK study for stricturing phenotype. Our study also identified new clinical findings such as low albumin as a risk factor for stricturing phenotype and lower hemoglobin as well as elevated ESR as risk factors for penetrating disease activity. Hypoalbuminemia, anemia, and elevated ESR are indicators of increased disease activity, which could suggest a more severe disease course.

Table 2.2: Comparison of the findings from our study (PEDSnet) with the RISK study in PCD. The results from the RISK study are from Table 2 in Kugathasan et al. (2017). The blank space in the table indicates missing values from each study.

	Stricturing behavior		**Penetrating behavior**	
	RISK* (n=913)	**PEDSnet (n=2,944)**	**RISK* (n=913)**	**PEDSnet (n=2,944)**
Age at dx in years	1.07 (0.97–1.17)	1.12 (1.08–1.16)	1.45 (1.17–1.80)	1.13 (1.09–1.17)
African American race	1.08 (0.52–2.22)	1.25 (0.91–1.71)	3.19 (1.39–7.31)	1.00 (0.71–1.41)
Isolated ileal location	1.60 (0.88–2.91)	—	1.23 (0.51–2.95)	—
ASCA IgA positive	1.69 (0.94–3.07)	—	2.68 (1.19–6.04)	—
Cbirl positive	2.30 (1.26–4.20)	—	3.01 (1.31–6.93)	—
Albumin	—	0.81 (0.67–0.99)	—	0.81 (0.65–1.01)
BMI				
Severe malnutrition	—	1.15 (0.75–1.74)	—	1.05 (0.67–1.64)
Moderate malnutrition	—	0.94 (0.65–1.35)	—	0.94 (0.62–1.44)
Normal	—	0.60 (0.46–0.77)	—	1.00 (0.75–1.34)
ESR	—	1.00 (1.00–1.01)	—	1.01 (1.00–1.01)
Hemoglobin	—	1.02 (0.95–1.09)	—	0.93 (0.87–0.99)
Hispanic race	—	1.49 (0.98–2.28)	—	1.29 (0.84–1.97)
Male	—	0.76 (0.61–0.95)	—	1.53 (1.21–1.94)

— indicates the data source does not have the risk factors.

Additionally, male sex was found to be a risk factor for stricturing (0.76, 95% confidence interval (CI) 0.61-0.95) and penetrating disease (1.53, 95% CI 1.21-1.94). This observation is in line with Mazor et al. (2011), who reported a significant association between male sex and PCD through multivariate logistic regression (odds ratio (OR) 2.60, 95% CI 1.17-5.75). It's worth noting

thatMazor et al. (2011) employed a multivariate logistic regression model distinct from the ODACoRH algorithm on the Fine-Gray competing risks model. Additionally, the ODACoRH algorithm revealed a correlation between low albumin levels and stricturing behavior (0.81, 95% CI 0.67-0.99). This finding aligns with the results of a study by Kori et al. (2022), which identified low albumin levels (OR, 3.97, 95% CI: 1.32-11.97) at diagnosis as a predictor of complicated disease in adulthood. The unexpected finding of normal BMI as a risk factor for the development of stricturing phenotype may be due to the exclusion of patients with severe malnutrition, who are more likely to present with strictures at diagnosis.

Evaluation of impacts of risk factors for the recurrence of OM while concurrently considering the subsequent TTI for OM study

We implemented the ODACoRH algorithm on the OM study across seven children's hospitals within the PDESnet. This application aimed to compare the results generated by the ODACoRH algorithm with those obtained through the pooled data analysis and the meta-analysis. The study cohort comprised 51,395 children under the age of 18 who had the OM and the initial TTI. Among this cohort, the primary event of interest was the recurrence of the OM (34.4%) and the competing event was the subsequent TTIs (10.0%). The time for the recurrence of the OM or the TTIs is recorded.

The ODACoRH algorithm successfully identified risk factors associated with the recurrence of the OMs, which were age at entry, year at entry, the medical history of cleft palate, and delays in speech and language acquisition. Notably, these findings aligned with those derived from the pooled data analysis. However, differences emerged when comparing the results with those from the meta-analysis, where the risk factor of male sex and the medical history of cystic fibrosis were misidentified in contrast to the pooled method. Additionally, the meta-analysis failed to recognize the significance of a risk factor related to the medical history of delays in speech and language acquisition. The results of the risk factors associated with the recurrence of the OM from the ODACoRH algorithm, the meta-analysis, and the pooled method are shown in the Supplementary Materials.

The validation process conducted across the PCD and OM real-world cases underscores the robustness and generalizability of the ODACoRH algorithm in handling diverse data with competing risks

and heterogeneity.

2.4. Discussion

The ODACoRH algorithm is a communication-efficient method that does not require sharing patient-level data. It is capable of accounting for the heterogeneity of baseline subdistribution hazard functions, resulting in more accurate results compared to models that do not account for heterogeneity.

From the extensive numerical experiments, we found out that the results from the meta-analysis method possess a larger bias than the ODACoRH algorithm especially when the event rates for multiple events are low. This is consistent with the findings of Duan et al. (2022) for the one-shot algorithm for a single type of time-to-event outcome (Duan et al., 2020a). In situations of low event rates, certain local sites may encounter either no events or very few occurrences, leading to significantly biased local-site estimates (Heinze and Schemper, 2001; King and Zeng, 2001). The meta-analysis, seen as a weighted average of individual site estimates, is severely impacted by biased estimates from sites with limited or no events. Outcomes in this study are defined using medical codes within the EHR system, which can be subject to misclassification and error. Recent literature has focused on methods to address these issues (Jing et al., 2025; Lu et al., 2024). We plan to report on such corrections in future work.

In contrast, the ODACoRH algorithm, as a likelihood-based approach, effectively mitigates this bias. This is attributed to the pooled data inherently containing a sufficient number of events, even in situations where events from specific sites are severely limited. By approximating the likelihood from the pooled data through the utilization of summary-level statistics, the ODACoRH algorithm consistently yields estimates with a smaller bias. For the relative efficiency of ODACoRH to the meta-analysis method, the asymptotic variance of the ODACoRH and meta-analysis are equivalence to the pooled method, when the number of patients per site increases and the number of sites is fixed (Jordan et al., 2019; Lin and Zeng, 2010).

The algorithm can handle multiple types of outcomes and terminal events, allowing for a comprehensive understanding of disease progression. However, it is not suitable for semi-competing risk

data, which involves multiple outcomes with both non-terminal and terminal events, such as time from MCI to AD/ADRD and time from MCI to death. We plan to develop distributed algorithms for semi-competing risk data in the future. Additionally, the ODACoRH algorithm can be extended to accommodate time-varying covariates, but this will affect the interpretation of subhazard ratios (Zhou et al., 2012).

One potential limitation of this study is that it only includes data up until August 2020, which includes the first six months of the SARS-CoV-2 pandemic in the United States. While the pandemic may have caused some delays in patient care, there is currently no evidence to suggest that COVID-19 infection worsens disease activity in patients with Crohn's disease (Bezzio et al., 2023). Therefore, the study period was kept through August 2020. Another limitation is missing data, particularly regarding disease location. The RISK study identified disease location as a risk factor for complicated disease behavior, but due to missingness in the dataset, this variable could not be analyzed in the current study. Lastly, PCD may consist of distinct subphenotypes, and the distribution of these subphenotypes could vary across sites. Recent literature highlights the use of transfer learning to identify subphenotypes and estimate their proportions within target populations (Wu et al., 2024a). We plan to apply similar approaches to characterize PCD subphenotypes and report our findings in future work.

2.5. Conclusions

The proposed ODACoRH is a communication-efficient algorithm to effectively handle between-site heterogeneity for the competing risks model. It enables to identification of the impact of the risk factors on multiple types of events. The validation of the ODACoRH algorithm was conducted through extensive numerical experiments and applied to two real-world cases concerning PCD and OM within the PEDSnet. The results of the ODACoRH algorithm consistently demonstrated superior performance compared to the meta-analysis, particularly when addressing rare event rates. These findings underscore the algorithm's robustness and efficacy in handling complex scenarios involving multiple types of events and heterogeneity, making it a valuable tool for advancing research in the field.

2.6. Code Availability

The R codes of the ODACoRH algorithms are available at https://github.com/Penncil/ODACoRH and through our R package `pda` at https://github.com/Penncil/pda.

CHAPTER 3

BIAS REDUCTION IN DISTRIBUTED INFERENCE FOR RARE EVENTS

3.1. Introduction

In a regular model indexed by parameters θ, it is well-known that the maximum likelihood estimates $\hat{\theta}$ can exhibit first-order asymptotic bias (Firth, 1993). Under rare event settings, this first-order bias can be of the order of $O(n_e^{-1})$, where n_e is the expected number of events (King and Zeng, 2001), leading to non-negligible biases. This increased bias is particularly significant in rare disease analyses, including pharmacovigilance and drug safety studies, where the maximum likelihood estimator $\hat{\theta}$ can approach infinity with zero-event data (Heinze and Schemper, 2002). For instance, quantifying risk factors for eosinophilic granulomatosis with polyangiitis is crucial due to its rapid progression and potential for severe damage to vital organs. However, it is challenging given its low incidence of approximately three to five cases per million annually (Gokhale et al., 2021), where the maximum likelihood estimates can be significantly affected by the first-order bias. To reduce this bias, a bias-corrected likelihood (Firth, 1993) is proposed by introducing a bias-reducing factor into the estimating equation of θ. This method has been widely used (Heinze and Schemper, 2002; Joshi et al., 2022) and consistently yields improved estimates in the rare event scenarios (Kosmidis and Firth, 2010, 2021; Wang, 2014).

The same bias issue arises in distributed learning with decentralized data when analyzing rare events. In recent years, distributed research networks have emerged as a fundamental tool in the study of rare diseases. For example, “Predictive Analytics via Networked Distributed Algorithms for Multi-System Diseases" is a national research initiative that seeks to leverage electronic health records across ten major national health systems within the OneFlorida+ Clinical Research Network (Hogan et al., 2022) to facilitate the early diagnosis and treatment of multi-system rare diseases. Due to practical and regulatory constraints, most distributed research networks do not permit patient-level data to be shared among sites, necessitating the use of distributed algorithms, or federated learning, which enables multi-center joint analysis by sharing only summary-level statistics

(Wang et al., 2017; Jordan et al., 2018; Duan et al., 2022; Li et al., 2024). While methods like the simple average estimator (Zhang et al., 2012) are popular, they have been primarily studied in scenarios where the probability of event occurrence is moderate. Research on their performance in rare event settings, where first-order bias can lead to inaccurate or infinite estimates, is limited. Additionally, to the best of our knowledge, bias-reduction techniques for distributed inference in scenarios with near-zero event probabilities have not been explored.

To tackle this challenge, we develop and study a novel distributed inference framework for bias correction. Our algorithm requires only one or two rounds of between-site communication. Specifically, by averaging the bias-corrected local estimators obtained from each site, we get the one-round distributed Firth estimator. Based on this, we further construct a novel surrogate likelihood to approximate the pooled data-based bias-corrected likelihood by utilizing the patient-level data from a lead site and summary-level statistics from other sites. The two-round distributed Firth estimator is obtained by maximizing the surrogate likelihood function. We situate our theoretical analysis within a novel context concerning distributed inference for rare events where n_e, the expected number of events per site, approaches infinity at a significantly slower rate as the site sample size increases, leading to a zero-approaching probability of event occurrence. The number of sites is fixed in the analysis.

To the best of our knowledge, this is the first rigorous investigation of systematic bias of distributed estimators under a rare event setting. The contribution of this paper is threefold:

- We proposed a communication-efficient distributed learning approach, which not only effectively eliminates the first- and second-order asymptotic biases in the estimates but also avoids iterative communication between sites.

- We clarified why the simple average estimator, which is widely used for decentralized data, performs poorly in the rare event setting by showing that its leading asymptotic bias terms are tied to the number of events with the order of $O(n_e^{-1})$.

- We showed that the proposed one-round distributed Firth estimator can reduce the leading

asymptotic bias to the order of $O(n_e^{-2})$. Additionally, one more round of communication can further reduce the leading asymptotic bias term to the order of $O(n_e^{-2.5})$.

We conducted simulation studies to support our theoretical findings. In addition, we applied our distributed Firth estimators to multi-center electronic health records data from the OneFlorida Clinical Research Network to identify the risk factors for the rare disease, eosinophilic granulomatosis with polyangiitis.

3.2. First-order bias of maximum likelihood estimation with distributed data

3.2.1. Setting

We focus on the asymptotic bias analysis of maximum likelihood estimation under the distributed inference framework. Specifically, from a fixed number of m sites, we collect in total $n \times m$ observations, where each site contributes n independently and identically distributed observations. The extension to unequal sample size across sites is straightforward, which is provided at the end of Section 3 and demonstrated in the real data analysis in Section 5. Data from m sites are decentralized, and individual-level data cannot be shared between sites, necessitating the use of distributed learning algorithms. Additionally, since between-site communication often involves human effort and can be time-consuming, algorithms that require only one or a few rounds of communication are preferred for distributed inference.

We consider an association study between a binary outcome Y and a set of covariates $X = (1, Z^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{R}^{p+1}$, and the dependence of Y on X is characterized by the generalized linear model with canonical link

$$f_{Y|X}(y|x) = c(y) \exp\left[\left\{yx^T\theta - h\left(x^T\theta\right)\right\}/\phi\right],$$

where $\theta \in \mathcal{R}^{p+1}$ is the unknown parameter of interest, $h(\cdot)$ is a known function, and ϕ is the known dispersion parameter. We write $\theta = (\alpha, \beta^T)^T$ with $\alpha \in \mathcal{R}$ being the intercept and $\beta \in \mathcal{R}^p$ being the regression coefficients, and the true value of θ is $\theta_0 = (\alpha_0, \beta_0^T)^T$. To estimate the parameter θ within the jth site, we can obtain the local maximum likelihood estimator $\hat{\theta}_j$ by maximizing the

local loglikelihood function $L_j(\theta) = n^{-1}\sum_{i=1}^{n} L(\theta; x_{ij}, y_{ij})$, where $L(\theta; x_{ij}, y_{ij}) = \log\left\{f_{Y|X}\left(y_{ij}|x_{ij}\right)\right\}$ is the loglikelihood for the ith patient in the jth site. When individual-level data from all sites can be pooled together, which is generally infeasible in practice, the global loglikelihood function based on the pooled data is $L(\theta) = (nm)^{-1}\sum_{j=1}^{m}\sum_{i=1}^{n} L(\theta; x_{ij}, y_{ij})$. By maximizing $L(\theta)$, we can obtain the maximum likelihood estimator $\hat{\theta}$ based on the pooled data.

In rare event analyses, when the event occurrence probability is extremely low, the asymptotic behavior of the maximum likelihood estimator can be severely affected by the number of cases and $\hat{\theta}$ can even be infinite (Owen, 2007). Among existing efforts of rare event analyses, although in a totally different context, Wang (2020) proposed a setting that can properly characterize this scenario. Specifically, to describe it in our context, as $n \to \infty$, we let the observed number of events n_{oj} within the jth site to satisfy

$$n_{oj}/n = \mathrm{pr}\left(Y = 1\right)\left\{1 + o(1)\right\} = E\left[h'(\alpha_0 + X^{\mathrm{T}}\beta_0)\right]\left\{1 + o\left(1\right)\right\} \to 0,$$

and

$$n_{oj} = n\mathrm{pr}\left(Y = 1\right)\left\{1 + o(1)\right\} = n_e\left\{1 + o\left(1\right)\right\} \to \infty,$$

for each $j = 1, \ldots, m$. In this setting, at each site, the rate at which n_{oj} (or n_e) goes to infinity is slower than that of n. This type of asymptotics is also considered in semi-supervised learning settings, see Chakrabortty and Cai (2018). We adopt this characterization of rare event settings in our investigation.

3.2.2. First-order Bias

In the following, we focus on logistic regression as an illustrating example, and the main results in this paper can also be applied to other models when the required assumptions are satisfied. With standard regularity conditions elaborated in the Supplementary Material, Lemma 1 characterizes the first-order bias and the mean square error of the maximum likelihood estimator $\hat{\theta}$ obtained from optimizing the pooled data-based likelihood function in rare event scenarios.

Lemma 3.2.1. *Under Assumptions 1-3 in the Supplementary Material, the bias of the maximum*

likelihood estimator $\hat{\theta}$ obtained from the pooled data satisfies $\|E(\hat{\theta}-\theta_0)\|_2 \leq C_1/(mn_e)$, where C_1 is a constant unrelated to n, m, and θ_0. Moreover, the mean squared error of the maximum likelihood estimator $\hat{\theta}$ satisfies $E\|(\hat{\theta}-\theta_0)\|_2^2 \leq C_2/(mn_e)$, where C_2 is a constant unrelated to n, m, and θ_0.

Lemma 1 shows that the performance of maximum likelihood estimation heavily depends on the sample size and the disease prevalence $pr\,(Y=1)$ explicitly through the expected number of events. Therefore, when estimating the unknown parameters in a single site, the first-order bias of the local maximum likelihood estimator $\hat{\theta}_j$, i.e., $n_e^{-1/2}$, is related to the expected number of events within the local site, which explains the potential of getting highly biased local estimates as a consequence of having a small n_e at a local site. By combining data across m sites, the pooled data-based maximum likelihood estimator $\hat{\theta}$ can benefit from a larger event number to improve the first-order bias rate to $(mn_e)^{-1/2}$. However, in the distributed inference framework, pooling patient-level data together to obtain the maximum likelihood estimate is usually not applicable. The simple average estimator $\bar{\theta}$ is widely used to approximate $\hat{\theta}$ by averaging all local maximum likelihood estimators, i.e., $\bar{\theta} = \sum_{j=1}^{m} \hat{\theta}_j/m$. However, the next lemma reveals that the leading bias term of the simple average estimator is still severely affected by n_e via n_e^{-1}.

Lemma 3.2.2. *Under Assumptions 1-3 in the Supplementary Material,the simple average estimator $\bar{\theta}$ satisfies*

$$E\|(\bar{\theta}-\theta_0)\|_2^2 \leq \frac{C_3}{mn_e} + \frac{C_4}{n_e^2},$$

with C_3 and C_4 are constants unrelated to m, n, and θ_0.

Lemma 2 decomposes the mean squared error of $\bar{\theta}$ into the summation of its variance and the squared bias, and it shows that the leading bias term of $\bar{\theta}$ is upper-bounded by a constant times n_e^{-1}. This indicates that, compared with the local estimator, the simple average estimator does not reduce the bias and it can still lead to inaccurate estimates. Therefore, new methods are needed and we introduce a novel distributed inference algorithm to reduce the bias and get a sharper upper bound of the estimation error in the next section.

3.3. Bias Correction using distributed Firth estimator

3.3.1. Firth estimator for pooled data

Within a single site, Firth (1993) modified the score function $S(\theta)$, the first derivative of the log-likelihood $L(\theta)$, to correct the first-order asymptotic bias of maximum likelihood estimators by adding $(2n)^{-1}\text{tr}\{I_j(\theta)\partial I_j(\theta)/\partial\theta_r\}$ to the rth element of $S(\theta)$, where $I_j(\cdot)$ is the observed information matrix obtained at the jth site and $\text{tr}(\cdot)$ denotes the trace of a matrix. For the logistic regression model, the modified score function is equivalent to the penalized loglikelihood function (Kosmidis and Firth, 2021)

$$L_{F,j}(\theta) = L_j(\theta) + (2n)^{-1}\log\det\{I_j(\theta)\}.$$

If the individual-level data could be shared across sites, with $I(\theta)$ being the Fisher information matrix of the pooled data, the pooled Firth-corrected loglikelihood function is

$$L_F(\theta) = L(\theta) + (2nm)^{-1}\log\det\{I(\theta)\}, \tag{3.1}$$

from which we obtain the bias-corrected estimator $\hat{\theta}_F$, which we will treat as a gold standard for comparison with other candidate distributed estimators. In practice, however, $\hat{\theta}_F$ is not available as individual-level data sharing is often prohibited due to privacy concerns. To approximate $\hat{\theta}_F$, in the next subsection, we develop a distributed inference algorithm that only requires sharing summary-level data between sites.

3.3.2. Distributed Firth Estimator

We denote the one-round distributed Firth estimator by $\tilde{\theta}^{(1)}$. To obtain $\tilde{\theta}^{(1)}$, we simply apply the divide-and-conquer strategy with two steps. The first step is to obtain the bias-corrected local estimator $\hat{\theta}_{F,j} = \arg\max L_{F,j}(\theta)$ by using local data within each site. The second step then calculates the average of the local estimators as

$$\tilde{\theta}^{(1)} = \textstyle\sum_{j=1}^{m}\hat{\theta}_{F,j}/m. \tag{3.2}$$

As we later show in Theorem 1, $\tilde{\theta}^{(1)}$ still suffers from a small n_e via $O\left(n_e^{-2}\right)$, the leading bias term. The estimator's robustness to the impacts of rare events can be further improved by conducting one more round of communication. Specifically, based on $\tilde{\theta}^{(1)}$, we approximate the pooled data-based Firth-corrected likelihood function in equation (3.1) using a surrogate function. We consider a practical scenario in distributed inference where we have a *lead site* whose individual-level data are accessible while other sites can only share summary-level statistics. Without loss of generality, we let site 1 be the lead site and construct the surrogate function of the pooled data-based Firth-corrected likelihood function as

$$\tilde{L}_F(\theta) = L_{F,1}(\theta) + \{\nabla L_F(\tilde{\theta}^{(1)}) - \nabla L_{F,1}(\tilde{\theta}^{(1)})\}^T\theta, \tag{3.3}$$

where ∇ is the first-order gradient with respect to θ. Note that $\nabla L_F(\tilde{\theta}^{(1)})$ in equation (3.3) can be calculated by using only gradient information from other sites, which are summary-level statistics. The details to obtain $\nabla L_F(\tilde{\theta}^{(1)})$ are in the Supplementary Materials. By finding the root of $\nabla\tilde{L}_F(\theta) = 0$, we obtain the two-round distributed Firth estimator $\tilde{\theta}^{(2)}$.

Our distributed algorithm is summarized in Algorithm 1. We use T as a controller of the communication rounds with $T = 1$ or $T = 2$ corresponding to $\tilde{\theta}^{(1)}$ or $\tilde{\theta}^{(2)}$ with one or two rounds of communications, respectively. In Algorithm 1, we use ∇^e to denote the eth-order gradient ($e = 1, 2, 3$). Only summary-level statistics, including ∇L_j, $\nabla^2 L_j$, and $\nabla^3 L_j$ for the jth site ($j = 1, \ldots, m$), are shared across sites. There is no requirement to share patient-level data.

The next result characterizes the mean squared error of our proposed distributed estimators.

Theorem 1. *Under Assumptions 1-3 in the Supplementary Material, the one-round distributed Firth estimator $\tilde{\theta}^{(1)}$ satisfies*

$$E\|(\tilde{\theta}^{(1)} - \theta_0)\|_2^2 \leq \frac{C_5}{mn_e} + \frac{C_6}{n_e^4},$$

where C_5 and C_6 are constants unrelated to m, n, and α_0. If one more round of communication is

Algorithm 2 Algorithm for the Distributed Firth Estimator

Initialization:
In the lead site ($j = 1$), compute $\hat{\theta}_{F,1}$ by maximizing $L_{F,1}(\theta)$.
for $j = 2$ to m **do**
    Compute $\hat{\theta}_{F,j}$ by maximizing $L_{F,j}(\theta)$.
    Transfer $\hat{\theta}_{F,j}$ to the lead site.
end for
In the lead site, calculate $\tilde{\theta}^{(1)}$ using Equation (3.2).
if $T = 1$ **then**
    Output: $\tilde{\theta}^{(1)}$ (one-round distributed Firth estimator).
else
    Broadcast $\tilde{\theta}^{(1)}$ to all sites.
    for $j = 2$ to m **do**
        Compute $\nabla^e L_j(\tilde{\theta}^{(1)})$ for $e = 1, 2, 3$, the eth-order derivative of $L_j(\theta)$ at $\tilde{\theta}^{(1)}$.
        Transfer $\nabla L_j(\tilde{\theta}^{(1)})$, $\nabla^2 L_j(\tilde{\theta}^{(1)})$, and $\nabla^3 L_j(\tilde{\theta}^{(1)})$ to the lead site.
    end for
    In the lead site, construct $\tilde{L}_F(\theta)$ using Equation (3.3).
    Compute $\tilde{\theta}^{(2)}$ by maximizing $\tilde{L}_F(\theta)$.
    Output: $\tilde{\theta}^{(2)}$ (two-round distributed Firth estimator).
end if

conducted, then the two-round distributed Firth estimator $\tilde{\theta}^{(2)}$ *satisfies*

$$E\|(\tilde{\theta}^{(2)} - \theta_0)\|_2^2 \leq \frac{C_7}{mn_e} + \frac{C_8}{n_e^5},$$

where C_7 *and* C_8 *are constants unrelated to* m*,* n*, and* α_0*.*

Theorem 1 shows that our one-round estimator $\tilde{\theta}^{(1)}$ has successfully reduced the rate of its leading bias term to be $O(n_e^{-2})$, a smaller term compared with that of the simple average estimator. Furthermore, by one more round of communication, $\tilde{\theta}^{(2)}$ further reduces the rate of the leading bias term to be $O(n_e^{-2.5})$, even sharper than that of the one-round distributed Firth estimator.

Remark 1. We can extend all the above results to the unequal sample size scenario and the order of the leading bias terms for the proposed method are not changed. The leading bias term for $\tilde{\theta}^{(1)}$ is related to the expected number of events within the smallest site and the leading bias term for $\tilde{\theta}^{(2)}$ is related to the expected number of events for the leading site. More details can be found in the Supplement Material.

3.4. Simulation Studies

We conduct simulations to demonstrate the performance of the proposed method. Logistic regression is considered to characterize the association between an exposure Z_1 and an outcome Y while adjusting for a confounding variable Z_2. We generate Y_{ij} from

$$\text{pr}(Y_{ij} = 1|Z_{ij1}, Z_{ij2}) = h(\alpha + Z_{ij1}\beta_1 + Z_{ij2}\beta_2),$$

where Z_{ij1} follows a Bernoulli distribution with probability p_1, Z_{ij2} follows a Bernoulli distribution with probability p_2. To simulate the distributed inference setting, we let $m = 10$ and $n = 1,000$. The methods under comparison include the simple average estimator $\bar{\theta}$ and our proposed estimators $\tilde{\theta}^{(1)}$ and $\tilde{\theta}^{(2)}$. To validate our theoretical results, we change the value of θ to simulate both the rare event setting and the general setting where we have common outcomes. Moreover, to mimic the subsequent real-world data analysis, the value of p_1 is also changed to get different levels of prevalence of the exposure. Therefore, the following four scenarios are considered:

1. both the outcome and the exposure are common;
2. the outcome is common and the exposure is rare;
3. the outcome is rare and the exposure is common;
4. both the outcome and exposure are rare.

For each scenario, the values of θ and p_1 are in Supplementary Material. Under each scenario, we replicate the analysis 200 times and report the results in Figure 3.1.

Since the pooled data-based bias-corrected estimator $\hat{\theta}_F$ is the gold standard in our setting, which is normally not available in practice though, we measure the relative bias of the estimates of β_1 obtained from different methods to $\hat{\theta}_F$ as a metric in method comparison. From Figure 3.1, we observe that the simple average estimator shows significant bias in all four scenarios, and the bias is amplified in rare outcome settings. The one-round and two-round distributed Firth estimators

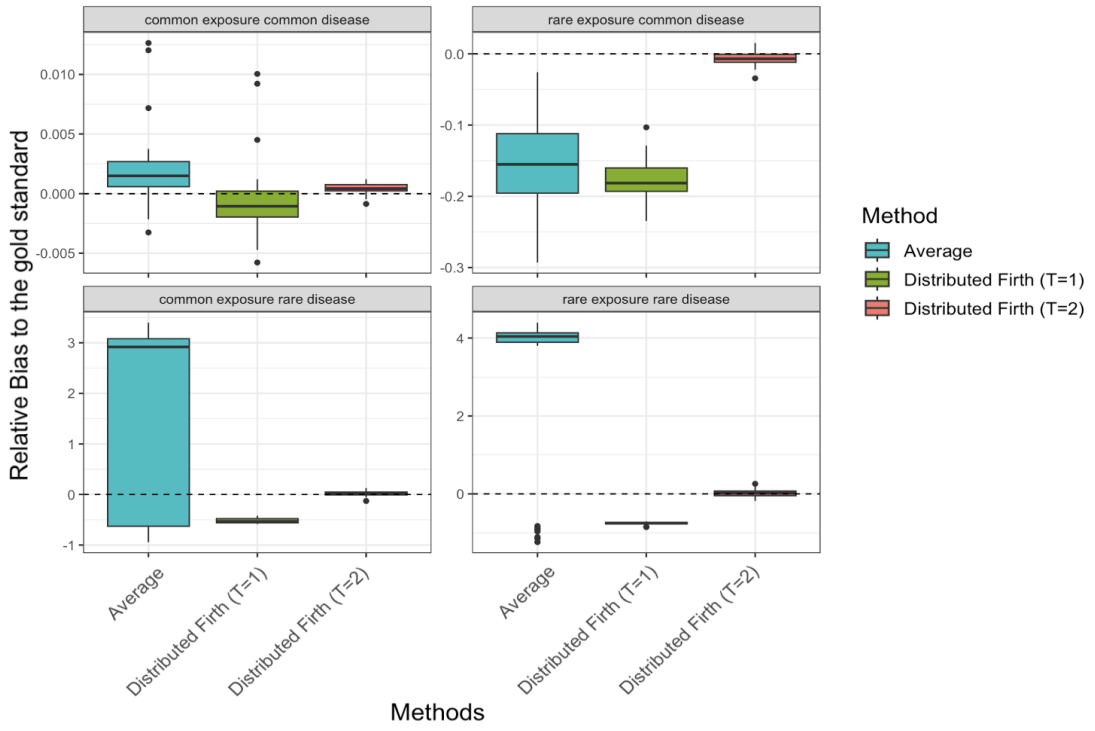

Figure 3.1: Boxplot of relative bias of simple average estimator and distributed Firth estimators to the gold standard estimator. The blue box represents the simple average estimator. The green box represents the one-round distributed Firth estimator. The red box represents the two-round distributed Firth estimator.

effectively eliminate the bias and achieve similar performance to the gold standard estimator in rare outcome settings. As for the rare exposure, it may affect the performance of both the average estimator and the one-round distributed Firth estimator, while the two-round distributed Firth estimator is robust to the rarity of the exposure and can still effectively correct the bias. To conclude, among three estimators, the two-round distributed Firth estimator can best approximate the gold standard estimator across all four scenarios and is the recommended estimator in distributed rare event analysis to obtain a bias-reduced estimator.

3.5. Data Analysis

We applied our approach to identify risk factors associated with eosinophilic granulomatosis with polyangiitis, a rare autoimmune vasculitis characterized by inflammation of small to medium-sized blood vessels (Wechsler et al., 2017). Electronic health records were extracted from eight clinical

sites in the OneFlorida+ Clinical Research Network. Our study period spans from December 2012 to July 2022. The study population included individuals who aged greater than or equal to 18 and without pre-existing diagnoses of the eosinophilic granulomatosis with polyangiitis before the study start date. Cases were defined as patients with diagnosis codes of eosinophilic granulomatosis with polyangiitis during the study period, while controls were defined as patients without any such diagnosis codes. We identified in total 129 cases, and $44,514$ controls were randomly selected from the eight clinical sites, resulting in a dataset where the overall prevalence of eosinophilic granulomatosis with polyangiitis was 0.28%. For each clinical site, the sample size is 3959, 4471, 5435, 7047, 3652, 6999, 7222, and 5858, respectively. The considered potential risk factors include age, sex, ethnicity/race, and medical history of asthma, glomerulonephritis, sinusitis, and alveolar hemorrhage within 18 months before the diagnosis of eosinophilic granulomatosis with polyangiitis.

Since the OneFlorida+ Clinical Research Network operates as a centralized data network, which enables us to access the pooled data, we get $\hat{\theta}_F$ as the gold standard. We then obtained the simple average estimator and the two-round distributed Firth estimator by using site 1 as the lead site. Comparisons are made between these two estimators and the gold standard estimator. Specifically, three methods provide consistent conclusions for most variables (see estimation results in the Supplementary Materials) except for the medical history of asthma and the history of glomerulonephritis. We present the estimated coefficients along with their corresponding 95% confidence intervals for these two variables in Figure 3.2. As for the medical history of glomerulonephritis, the simple average method yields an estimated effect size of 21.87 in odds ratio with a 95% confidence interval of $[5.64, 84.77]$, while the gold standard pooled analysis provides the odds ratio 2.28, with 95% confidence interval $[0.80, 6.49]$. On the other hand, our two-round distributed Firth estimator (odds ratio 2.64 with 95% confidence interval $[0.95, 7.39]$) provides a consistent result as the gold standard. We observed a similar situation for the medical history of asthma. The discrepancy between the results obtained by the simple average method and the gold standard analysis could be caused by the low prevalence of the disease, as well as the low prevalence of medical history for asthma (1.2%) and glomerulonephritis (1.3%). In this context, as demonstrated in simulations, the bias of the simple average method becomes substantial while our estimator can provide a good

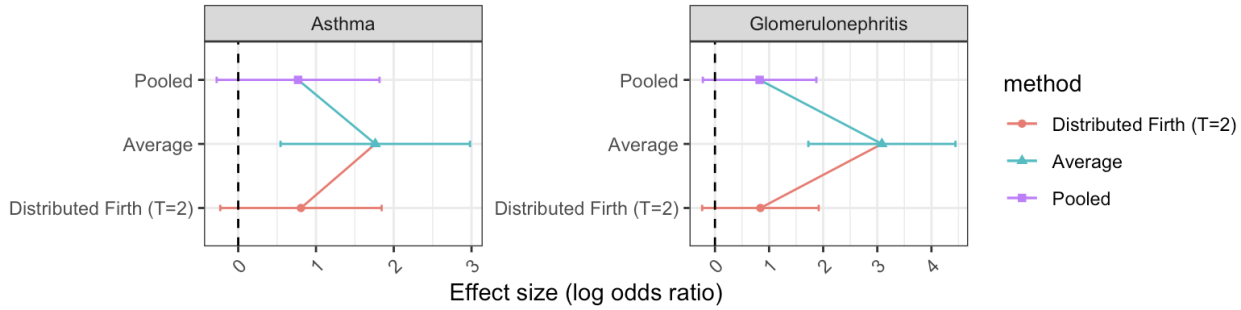

Figure 3.2: Boxplot of estimates and 95 percentage confidence interval of the simple average estimator, distributed Firth estimators, and the pooled estimator.

approximate to the gold standard estimator.

3.6. Discussion

We have presented a novel distributed learning framework aimed at reducing estimation bias, which is particularly useful when dealing with rare outcomes. By only one or two rounds of communication, our framework can yield estimates with significantly improved accuracy by reducing the first-order asymptotic bias. The theoretical performance of the proposed framework is demonstrated via logistic regression in a setting with a homogeneous rare outcome prevalence across sites. However, our algorithm can also be extended to other models used in modeling the rare outcomes, such as the Cox regression model for survival outcomes (Cox, 1972). Moreover, our method can be applied to a more general setting where different sites can preserve different incidences of rare events. For example, when the collaboration is in the global-level distributed research networks, the incidence of eosinophilic granulomatosis with polyangiitis can differ between countries (Jakes et al., 2021). Our future research will be focused on these topics.

CHAPTER 4

NEGATIVE CONTROL-CALIBRATED DIFFERENCE-IN-DIFFERENCE ANALYSES: ADDRESSING UNMEASURED CONFOUNDING IN REAL-WORLD DATA

4.1. Introduction

Difference-in-differences (DiD) analysis is a statistical method used to estimate causal effects by comparing changes in outcomes over time between an intervention group and a control group (Callaway and Sant'Anna, 2021; Abadie, 2005). In political science and biomedical research, DiD analysis has been used to assess the impact of interventions (Dimick and Ryan, 2014; Wing et al., 2018; Whittaker et al., 2016; Morris et al., 2014) or policy changes (Burden et al., 2014; Hansen, 2013; Neureiter, 2019; Malesky et al., 2014). For example, the DiD analysis assesses the racial/ethnic differences in health insurance after Medicaid expansion for patients with diabetes (Angier et al., 2019).

The rise of real-world evidence (RWE) has created significant opportunities for the application of DiD analysis. RWE, which often utilizes large-scale real-world data, offers valuable insights into the effectiveness and safety of interventions outside of controlled experimental settings (Sherman et al., 2016; U.S. Food and Drug Administration, 2021). This includes data from insurance claims, electronic health records, digital mobile health, and more. For instance, the Observational Health Data Sciences and Informatics (OHDSI) initiative captures a broader spectrum of patient populations and inter-country clinical practices, enhancing the findings' external validity (Hripcsak et al., 2015). DiD analysis, with its quasi-experimental design, is particularly advantageous for RWE, providing high-quality evidence for clinical decision-making.

A critical challenge in DiD analysis is the reliance on the parallel trends assumption. This assumption states that, in the absence of intervention, the intervention and control groups would, on average, have followed a parallel trajectory over time. This means that the effect of the unmeasured confounders on the outcomes is time-fixed, or not changing over time. However, a frequent issue arises when the effect of the unmeasured confounders varies with time (Pazzagli et al., 2018;

Streeter et al., 2017). Such variation can lead to diverging trends between the control and intervention groups and, if not accounted for, can introduce a systematic bias to the estimated effects of the intervention. Consequently, failing to account for the time-varying effect of unmeasured confounders and resulting violations of parallel trends may affect the reliability and generalizability of DiD analysis.

Many statistical methods have been proposed to solve the challenge of assessing and providing a statistical solution for violating the parallel trends assumption (Richardson et al., 2023; Gibson and Zimmerman, 2021; Ye et al., 2023). One approach is to match the measured confounders during the pre-intervention period before conducting the DiD analysis, aiming to simulate an environment where the parallel trends assumption is more likely to hold (Burden et al., 2014; Stuart et al., 2014; Imai et al., 2023). With empirical evaluation from numerical experiments, such a method mitigates the effect of the unmeasured confounders (Ham and Miratrix, 2024). However, even after matching, it is still undecided whether the parallel trends assumption is statistically valid. Performing valid hypothesis testing in this context is essential, as it would help evaluate the robustness of the DiD analysis. Furthermore, another significant challenge arises when attempting to address or correct systematic biases when the assumption does not hold.

In recent years, negative control outcome (NCO) experiments (Schuemie et al., 2018, 2014), which assume no intervention effect on the NCO, have been used to calibrate the systematic confounding biases related to the intervention effect. The NCO experiments are generally limited to outcomes measured after the intervention period. However, applying the NCO assumption to both pre-intervention and post-intervention outcomes offers a more accurate calibration of confounding biases. Thus, developing methods to incorporate NCOs in DiD analysis to correct systematic biases is crucial for ensuring the robustness of causal inferences drawn from DiD studies.

We propose a paradigm-shifting novel framework for DiD analysis, termed negative control-calibrated difference-in-difference method (NC-DiD), including a hypothesis testing procedure for possible violation of the parallel trends assumption and a data-driven calibration process utilizing the NCO experiments. This calibration procedure is structured to mitigate potential biases resulting from

the violations of the parallel trends, thereby preserving the integrity of causal conclusions.

We conducted extensive numerical experiments on synthetic data and applied our method to a study examining racial/ethnic differences in health outcomes following COVID-19 infection. Our research aimed to assess whether minority racial and ethnic groups experienced worse long-term outcomes after COVID-19 compared to Non-Hispanic White (NHW) groups. Previous research by (Asch et al., 2021a) showed that Non-Hispanic Black (NHB) patients hospitalized with COVID-19 had higher 30-day mortality compared to NHW patients, but the long-term effects of COVID-19 on NHB and other minority groups remain unclear. In our study, we evaluated racial and ethnic differences in post-COVID outcomes among 15,373 patients under the age of 21 from eight children's hospitals. The numerical experiments and the real-world study demonstrate that NC-DiD effectively detects and corrects systematic bias, offering a more rigorous and flexible methodology for researchers across various disciplines, including biomedical and political sciences.

4.2. Results

Overview of the Proposed Calibration Method

Traditional DiD models rely on the parallel trends assumption, which posits that differences in outcomes between treated and control groups would remain consistent over time in the absence of an intervention. However, this assumption is often violated due to time-varying effects of unmeasured confounders, introducing systematic bias. This bias undermines the accuracy of causal effect estimates, rendering the intervention's true effect unidentifiable. To address this challenge, we developed a negative control-calibrated DiD (NC-DiD) approach, incorporating a three-step calibration process to improve the robustness of causal inference.

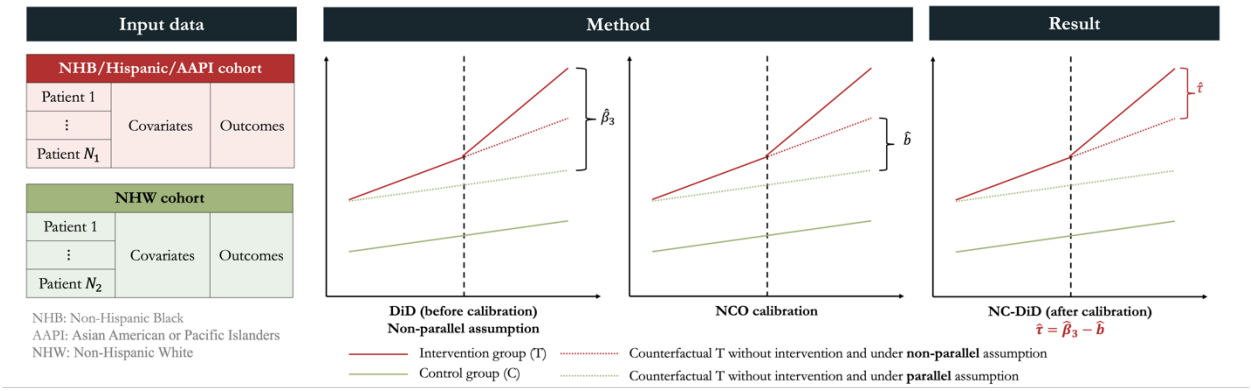

Figure 4.1: Workflow for the proposed NC-DiD method, consisting of three steps. First, implement the DiD method without calibration to obtain the estimated intervention effect, $\hat{\beta}_3$. Second, conduct negative control outcome experiments to estimate the systematic bias, b. Third, calibrate the process to obtain the calibrated intervention effect, $\hat{\tau}$.

As shown in the Figure 4.1, the first step involves implementing a standard DiD analysis, estimating the intervention effect while accounting for measured confounders. Next, negative control outcome (NCO) experiments are conducted under the assumption that the intervention does not affect these control outcomes. By applying the DiD model to each NCO, systematic bias is estimated through an aggregation process. We proposed two ways to aggregate the information from each NCO to identify the systematic bias. The empirical posterior mean approach combines information from all NCOs by taking their weighted average, consistent with empirical Bayes principles where we assume the biases follow a normal distribution with unknown parameters. This approach is optimal when all NCOs are valid and follow our modeling assumptions. In contrast, the median calibration approach uses the median of the estimated biases across all NCOs, providing robustness against invalid NCOs.

This systematic bias estimate from NCOs also enables testing the validity of the parallel trends assumption. Finally, the intervention effect is calibrated by removing the estimated bias, providing a corrected and more reliable estimate of the intervention's true impact. This framework offers a systematic and robust approach to addressing bias, with theoretical guarantees detailed in the Supplement.

Synthetic data numerical experiments

Synthetic data was used to evaluate the proposed calibration method through extensive simulation studies. We first focused on the hypothesis testing for the parallel trends assumption. The significance level for the two-sided statistical test of the parallel trends assumption was set to 0.05. We simulated 40 NCOs and computed the empirical type-I error rate of the proposed test across varying sample sizes —1000, 2000, and 3000.

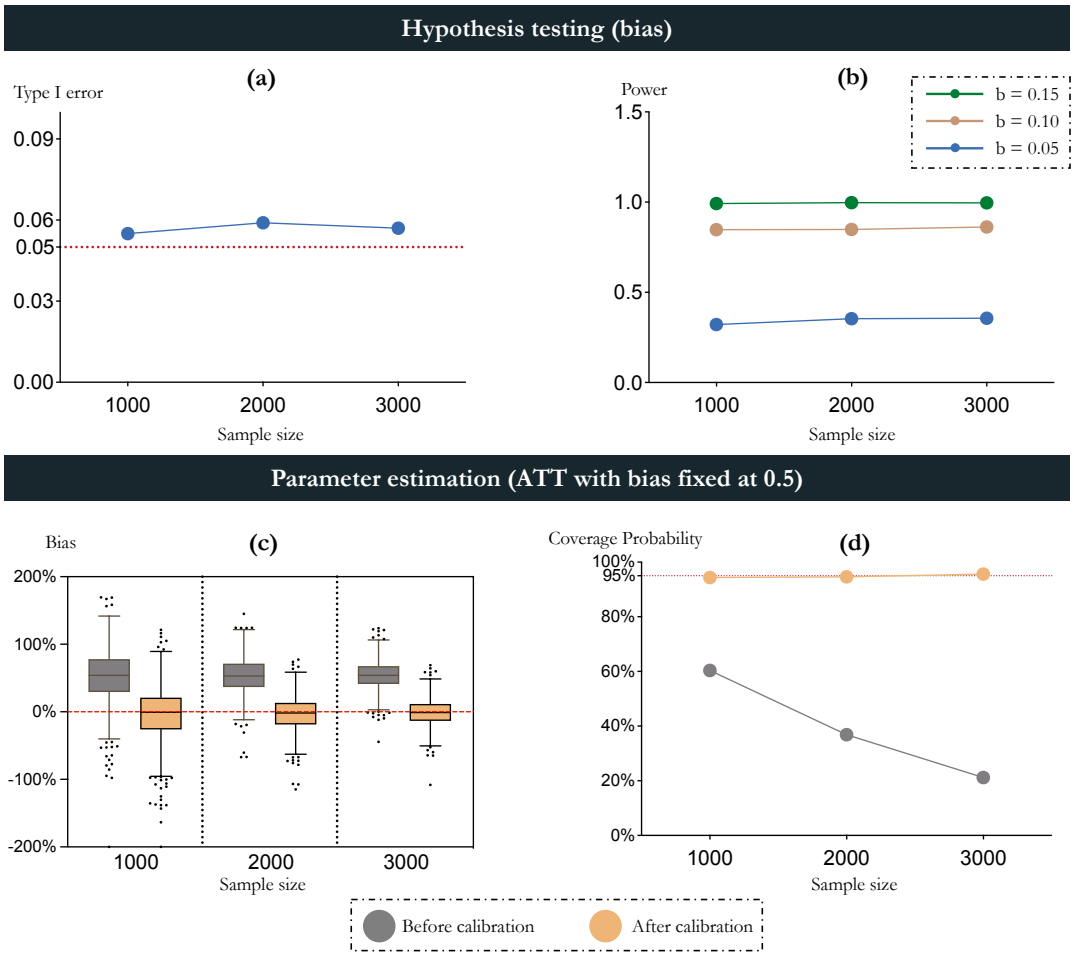

Figure 4.2: Results from 1,000 synthetic datasets. (a) Illustration of the Type-I error rate from the proposed NC-DiD method. (b) Statistical power for systematic bias values of $b = 0.05, 0.1, 0.15$. (c) Relative bias as the sample size increases from 1,000 to 3,000. (d) Coverage probability comparing the proposed NC-DiD method to the baseline DiD method as sample size increases from 1,000 to 3,000.

In Figure 4.2(a), the results show that, if the parallel trends assumption holds (b=0), the type-I error rate of the proposed test of parallel trends remains stable around 0.05 as the sample size goes larger, specifically, 0.055 for sample size 1000, 0.059 for sample size 2000, and 0.057 for sample size 3000. Figure 4.2(b) illustrates an increase in statistical power for rejecting the null hypothesis of parallel trends as the true value of b increases from 0.05 to 0.15. When b is increased to 0.1, the statistical power for sample size 1000 is 84.8%.

We further examined the impact of b on the estimation of the average treatment effect on the treated (ATT) in the post-intervention period using the DiD model when the parallel trends assumption was violated. We evaluated our proposed NC-DiD method, showing the relative bias and coverage probability for a 95% confidence interval of ATT. The relative bias for estimating the ATT was calculated as the ratio of the estimation error to the true ATT value. The coverage probability is defined as the proportion of times that a confidence interval contains the true parameter. In addition, we reported the mean square error (MSE), which was calculated as the average of squared differences between the estimated treatment effects and the true treatment effect across 1000 numerical experiments.

With a true value of the ATT set at -1 and b at 0.5, results from Figure 4.2(c) indicate a relative bias for the DiD model before calibration of 53.0%. This bias of the DiD model after the proposed calibration method was reduced to 2.6%. Furthermore, Figure 4.2(d) shows that after the calibration, the coverage probability was 95.6% near the nominal level of 95%, whereas without calibration, it was only 21.2% when the sample size was 3000. We evaluated the mean squared error (MSE) both before and after NCO calibration, with comprehensive results presented in Section C.3 across all simulation experiments. The results demonstrate substantial reductions in MSE following calibration. For instance, at a sample size of 3,000, the MSE for our proposed calibrated method was 0.039, compared to 0.331 for the baseline uncalibrated method, representing an 88% reduction in overall estimation error. This improvement in MSE, alongside the bias reduction, confirms that our calibration approach enhances overall estimation accuracy rather than simply shifting the estimation problem.

Handling Scenarios with Invalid Negative Control Outcomes (NCOs)

We evaluated the performance of the proposed calibration approach when incorporating partially invalid NCOs ((i.e., those not having a true treatment effect of zero). Figure 4.3 presents the results of the calibrated approach using both the empirical posterior mean and the median. The findings indicate that calibration based on the median outperforms the empirical posterior mean. Additionally, both the empirical posterior mean and the median methods demonstrate superior performance compared to the method without any calibration. For example, with a sample size of 1,000, the systematic bias using the median is -3.5%, compared to -19.8% with the uncalibrated method, which exhibits approximately 50% bias relative to the true value. Therefore, we recommend using the median-based calibration when the selected NCOs are not fully validated.

In summary, our simulations demonstrate that the proposed calibration method effectively maintains the type-I error rate across varying sample sizes and substantially reduces the relative bias caused by time-varying effects of unmeasured confounders in DiD models. This improvement in model accuracy is essential for ensuring reliable estimation of treatment effects, particularly in scenarios where the parallel trends assumption is violated.

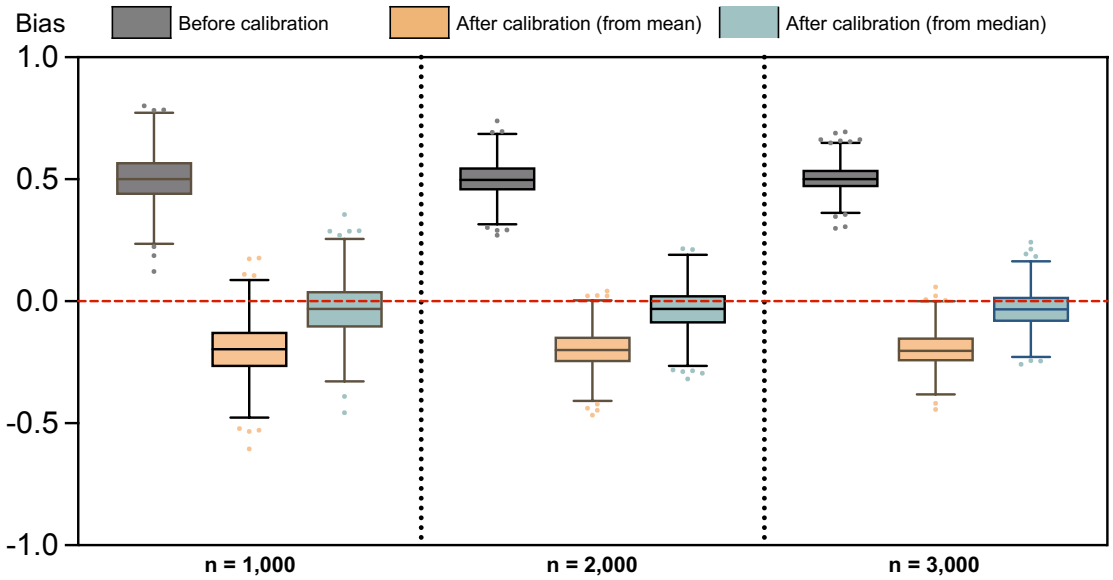

Figure 4.3: Results from 1,000 synthetic datasets under the scenario of partial validity of negative control outcomes (NCOs), comparing the calibration methods based on the empirical posterior mean and the median for estimating systematic bias.

Application to Racial/ethnical differences during the post-acute phase of COVID-19

Post-acute sequelae of SARS-CoV-2 infection (PASC), or long COVID, includes symptoms like fatigue, shortness of breath, and cognitive dysfunction persisting beyond 28 days after the initial infection. PEDSnet, a national pediatric health learning system provides us with an opportunity to understand, learn, and predict the PASC. Currently, PEDSnet includes data from over two million children from eight children's hospitals across the United States, including the Children's Hospital of Philadelphia and Stanford Children's Health. A detailed list of these hospitals and their corresponding abbreviations is provided in Supplementary Materials (Section C.1).

We investigate whether the infection of SARS-CoV-2 increases racial/ethnic differences in the PASC symptoms and conditions for minor race/ethnicity population. We aim to quantify the racial/ethnic differences attributable to COVID-19 infection. We defined the composite outcome as the presence of any healthcare visits related to 24 predefined PASC symptoms and conditions. The diagnosis code sets of the PASC symptoms and conditions are defined in Section C.2 of Supplement Appendix. We grouped the 24 PASC symptoms and conditions into systematic (conditions) and syndromic (symptoms) categories (Razzaghi et al., 2024). In addition, we used sixty-two NCOs with names of these NCOs detailed in Section C.3. We defined documented moderate to severe COVID-19 infection as cases where patients had positive PCR/antigen tests or a COVID-19 diagnosis, accompanied by moderately severe conditions such as gastroenteritis, dehydration, and pneumonia, or severe conditions requiring ICU admission or mechanical ventilation (Forrest et al., 2022). To illustrate our proposed method, we utilize a dataset including any patients under the age of 21 who had at least one documented moderate or severe COVID-19 infection. A total of 15,373 children and adolescents were included.

Within the cohort, 45.9% are female. We define the racial/ethnic groups as follows: NHW, NHB, Asian American or Pacific Islanders (AAPI), and Hispanic. The cohort composition was 53.7% NHW, 19.7% NHB, 4.5% AAPI, and 22.0% Hispanic. The index date for COVID-19 patients, which divides the two time periods, was defined by the first visit related to the documented COVID-19 infection between March 2020 and September 2023. We measured the three composite outcomes

for two time periods, the pre-infection period (28 days to 179 days before the index date), and the post-infection period (28 days to 179 days after the index date). We applied our proposed method by estimating the racial/ethnic difference attributable to COVID-19 infection between NHW and the three other racial/ethnic groups.

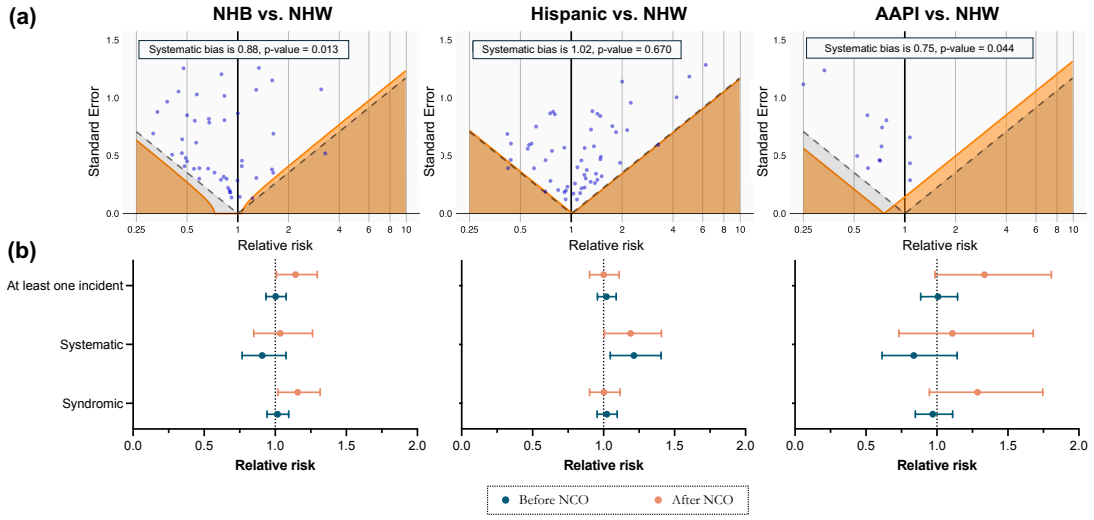

Figure 4.4: Results for the scientific question of investigating the racial/ethnic differences after the COVID-19 infection by the proposed calibration method via the empirical posterior mean. (a) showed the systematic bias in RR scale of the NHB, Hispanic, and AAPI compared with NHW. (b) demonstrated the results of the NC-DiD.

From Figure 4.4a, we observe that the test of parallel trends assumption for NHB versus NHW was rejected (p-value 0.013), and the estimated systematic bias was 0.88 on RR scale from the unmeasured confounding variables. Similarly, the estimated systematic bias on RR scale for AAPI versus NHW and Hispanic versus NHW were 0.75 (two-sided test p-value 0.044) and 1.02 (p-value =0.67), respectively.

We also show the results from the DiD method (before calibration) and the proposed method (after calibration) in Figure 4.4b. Among the AAPI group, we observe significant evidence of racial/ethnic differences due to COVID-19 infection after calibration. Interestingly, the results before calibration show moderate, though not significant, evidence for a difference in the opposite direction. After calibration, the RR for the prevalence of any visit with PASC symptoms or conditions after COVID-

19 infection comparing AAPI to NHW is 1.33 (95% confidence interval (CI) 0.98 to 1.81). Results before calibration showed RR in 1.01 (95% CI 0.89 to 1.15). Similarly, for any visits with syndromic PASC category after COVID-19 infection, the calibrated RR is 1.29 (95% CI 0.95 to 1.75) whereas for visits with systematic PASC category the calibrated RR 1.11 (95% CI 0.73 to 1.68). On the other hand, before calibration, comparing AAPI with the NHW group, the RR for visits with syndromic PASC category is 0.97 (95% CI 0.85 to 1.11) whereas the RR for visits with systematic PASC category 0.84 (95% CI 0.61 to 1.14).

Comparing the NHB and NHW groups, we found evidence of racial/ethnic differences for any visits to PASC symptoms and conditions after the COVID-19 infection by using the proposed calibration method. Specifically, the RR for the prevalence of any visits with PASC symptoms and conditions after the COVID-19 infection is 1.14 for NHB compared to NHW (95% CI 1.01 to 1.30). The RR in the syndromic PASC category for NHB compared with NHW is 1.16 (95% CI 1.02 to 1.32). No evidence was found for the risk of systematic PASC between NWB and NHW (RR 1.04, 95% CI 0.85 to 1.26). Comparing the Hispanic group with the NHW group, no evidence was found in the prevalence of any visits with PASC symptoms and conditions after COVID-19 infection (RR 1.00, 95% CI 0.90 to 1.11). However, the Hispanic group has a greater increase in the prevalence of the systematic PASC category after COVID-19 infection (RR 1.19, 95% CI 1.01 to 1.41).

4.3. Discussion

We present a novel framework that directly addresses the significant challenge in DiD analyses, that is, the parallel trends assumption may not hold due to the impact of unmeasured confounders with time-varying effects. The conventional DiD approach relies on the assumption that the difference between the intervention and control groups would have remained constant over time in the absence of an intervention. However, this assumption is frequently violated in practical scenarios, particularly in complex fields such as public health and policy research, where external factors can differ significantly between groups and over time. Through rigorous testing and calibration methods incorporated in our NC-DiD framework, we demonstrate that our approach not only identifies when the parallel trends assumption does not hold but also adjusts for this, providing a more accurate

estimation of causal effects.

While our NC-DiD approach shares mathematical relationships with alternative formulations, our framework offers distinct methodological advantages in RWD contexts. Our approach is particularly valuable when researchers need to perform hypothesis testing on the estimated bias from NCOs to evaluate the parallel trends assumption—a capability that would be bypassed in a single-step approach that simply treats NCOs as control groups during the post-intervention period. Additionally, our method provides a structured calibration procedure that fully leverages available pre-intervention data, enhancing estimation accuracy. Furthermore, our approach extends naturally to settings with multiple NCOs, where researchers can systematically integrate information from several negative controls to improve bias estimation through our framework. In these scenarios, the explicit calibration procedure provides both analytical transparency and interpretability advantages that enhance the method's utility for researchers working with complex RWD.

Our simulation studies support the effectiveness of the proposed framework in maintaining statistical power and reducing bias under various conditions, illustrating the robustness and adaptability of our approach. The practical relevance of our methodology is highlighted through its application to the study of racial/ethnic differences in PASC. Our results indicate significant differences in how racial groups are impacted after the COVID-19 infection, which traditional DiD approaches might have estimated incorrectly. By applying the NC-DiD method, we can offer more reliable insights that are crucial for developing targeted health policies such as equitable distribution of medical supplies and interventions such as community health outreach aimed at reducing these differences.

A notable innovation of our framework is its robust capability to handle partially invalid NCOs. In real-world scenarios, some NCOs may exhibit treatment effects, introducing additional bias if treated as entirely valid. Our framework mitigates this challenge by leveraging a robust calibration approach using the median effect of the NCOs rather than the empirical posterior mean. Our simulations demonstrate that this median-based calibration significantly reduces bias in the presence of invalid NCOs, ensuring reliable estimation of treatment effects even when not all NCOs meet strict validity criteria. This robustness is especially critical in complex observational studies where perfect NCO

selection may not always be achievable.

The robustness of our findings is supported by sensitivity analyses conducted across a subset of sixty-two negative control outcomes manually reviewed by clinicians (Section C.3). By comparing results between the full set and the subset NCOs, we demonstrate that our conclusions remain consistent regardless of which specific controls are included in the analysis. This consistency across different NCO configurations provides evidence that our method produces reliable causal estimates that are not dependent on particular control selection choices. Such sensitivity analyses should be considered standard practice when implementing the NC-DiD framework to ensure valid inference in the presence of potentially imperfect negative controls.

Looking ahead, we envision several promising future applications and extensions of NC-DiD within digital medicine. First, the framework can be integrated into healthcare systems' analytic pipelines to enable continuous monitoring of healthcare access across diverse populations. By implementing NC-DiD within existing clinical data warehouses, healthcare organizations could establish automated surveillance systems that detect differences in care access, quality, and outcomes, particularly following major healthcare disruptions or policy changes. Second, we anticipate that NC-DiD could be incorporated into digital clinical decision support systems to provide clinicians with more accurate estimates of treatment effects for specific patient subgroups. By adjusting for unmeasured confounding that varies across time and populations, NC-DiD could help personalize treatment recommendations based on more reliable effect estimates derived from an institution's own EHR data. Third, the methodology could be extended for real-time monitoring of public health interventions using streaming digital health data. As healthcare systems increasingly capture data continuously through digital interfaces, NC-DiD could be modified to provide progressive updates to effect estimates as new data become available, enabling more agile public health responses.

Our study has several limitations that warrant future research. First, the estimation of variance (and therefore construction of confidence intervals and hypothesis tests) for the calibrated estimates may be compromised by potential correlations among multiple NCOs. Nonetheless, our simulation results demonstrate that our model, even without addressing this correlation, effectively controls

the type I error rate and preserves the coverage probability of confidence intervals. Second, intervention misclassification could further impact the DiD model's accuracy. Recent studies36 propose calibrating this misclassification by specifying a range for the misclassification rate. We plan to investigate and adjust for this bias in future reports. Furthermore, we assume that the effects of each negative control outcome (NCO) share a common prior distribution, modeled as Gaussian. In the context of hypothesis testing, this assumption has proven empirically effective in accounting for heterogeneity among NCO effects. Future work could explore relaxing this parametric assumption by adopting non-parametric or semi-parametric approaches to model the distribution of NCO effects, potentially improving robustness when NCO effects exhibit non-Gaussian patterns. Such extensions might leverage techniques like kernel density estimation or Bayesian non-parametric methods to better capture the true underlying distribution of systematic biases across different outcomes in complex healthcare datasets.

In conclusion, this framework not only advances statistical methodology but serves as a practical tool for healthcare systems striving to utilize their digital repositories to improve patient outcomes across diverse populations. As healthcare systems expand their digital capabilities, NC-DiD enables more reliable evidence generation from observational data to guide clinical practice and health policy. Its ability to detect and adjust for time-varying unmeasured confounding transforms how researchers can leverage existing digital health infrastructure to produce trustworthy insights for clinical decision-making and health equity initiatives.

4.4. Methods

Notation

We consider an observational study conducted across two time periods denoted by T, where $T = 0$ denotes the pre-intervention period and $T = 1$ denotes the post-intervention period. Let Z_t represent the treatment assignment indicator with the restriction that treatment is assigned only at time $T = 1$ for the treated group, and $Z_T = 0$ for all units at $T = 0$. Additionally, we introduce the indicator A to distinguish the treated group from the control group at $T = 0$, defined such that $Z_T = A \times T$, where $A = 1$ if and only if $Z_1 = 1$.

We use the following counterfactual notation: $Y_t(z)$ denotes the potential outcome given treatment $Z = z$ at time $T = t$. Our target estimand is defined as:

$$\tau = \frac{E[Y_1(1) \mid A = 1]}{E[Y_1(0) \mid A = 1]}.$$

which can be interpreted as the average treatment effect on the treated (ATT) in terms of the RR. The primary distinction between the average treatment effect (ATE) and ATT lies in their target populations. The ATE is concerned with the overall population, whereas the ATT focuses specifically on those who receive the intervention. As our research aims to evaluate whether minority racial/ethnic populations experience worse long-term outcomes after COVID-19 infection, our proposed method focuses on the ATT. However, the method can also be applied to estimate the ATE. We use the RR because the outcome is binary, and RR is collapsible (Greenland et al., 1999; Whitcomb and Naimi, 2021). This framework can potentially be adapted for other target estimands, such as the odds ratio and risk difference, with corresponding adjustments.

Details of the proposed calibration method

The traditional difference-in-differences (DiD) model requires a parallel trends assumption, which states that the average difference in the outcomes between the intervention and control groups would remain constant over the two time periods if the treatment had not been applied. Typically, the assumption of parallel trends is violated, often when the effect of the unmeasured confounders is time-varying. This introduces a systematic bias, denoted b_0, to the model. Formally, the systematic bias can be quantified by

$$b_0 = \log\left(\frac{\mathbb{E}(Y_1(0) \mid A = 1)}{\mathbb{E}(Y_1(0) \mid A = 0)}\right) - \log\left(\frac{\mathbb{E}(Y_0(0) \mid A = 1)}{\mathbb{E}(Y_0(0) \mid A = 0)}\right). \tag{4.1}$$

Note that $b_0 = 0$ represents that the parallel trends condition holds for the outcome of interest. We refer to b_0 as the unmeasured confounder bias for the outcome of interest. Similarly, we can define the unmeasured confounder bias b_j for the jth NCO. Specifically,

$$b_j = \log\left(\frac{\mathbb{E}(W_j(0) \mid A=1)}{\mathbb{E}(W_j(0) \mid A=0)}\right) - \log\left(\frac{\mathbb{E}(W_j(0) \mid A=1)}{\mathbb{E}(W_j(0) \mid A=0)}\right). \tag{4.2}$$

As expected, the unmeasured confounder bias for different outcomes should be different. To account for the heterogeneity, we assume that these biases $b_0, b_1, \ldots, b_M \sim \mathcal{N}(b, \sigma^2)$, that is, they follow the normal distribution with mean b and variance σ^2. The parameter $b = 0$ can be interpreted as that the parallel trend assumption holds essentially, and nonzero b_js occur by some random error. The mean of the control counterfactual for the treated group during the after-intervention period is not generally identifiable for unmeasured confounding bias b. Hence the target estimand τ is not identifiable, either.

To mitigate the systematic bias, we proposed the following calibration procedure, as illustrated in Figure 4.1, consisting of three steps. In the first step, we adjusted the DiD model for measured confounders X, using a logistic regression model to estimate the propensity score given the measured confounders during the pre-intervention period. To implement the DiD method, we apply the log-linear model to the matched cohort:

$$\log\left(\mathbb{E}(Y \mid A, T)\right) = \beta_0 + \beta_1 A + \beta_2 T + \beta_3 AT. \tag{4.3}$$

Here, β_0 is a constant, and β_1, β_2, and β_3 are coefficients of A, T, and their interaction, respectively. Notably, β_3 represents the biased intervention effect in risk ratio (RR), which is related to the true intervention effect τ, that is,

$$\log(\tau) = \beta_3 - b_0.$$

The estimated $\hat{\beta}_3$ is derived from this model. In the second step, we repeat this procedure with M NCOs, under the assumption that the intervention has no effect on these outcomes. The intuition behind our test is straightforward: if the parallel trends assumption holds, then applying our DiD model to negative control outcomes should yield estimates centered around zero. Systematic deviation from zero would indicate violation of this critical assumption. Specifically, we fit the DiD model for the mth NCO, W_m ($m = 1, \ldots, M$):

$$\log\left(\mathbb{E}(W_m \mid A, T)\right) = \beta_{0,m} + \beta_{1,m}A + \beta_{2,m}T + \beta_{3,m}AT. \tag{4.4}$$

Here, $\beta_{3,m}$ is the coefficient of the interaction term, representing the systematic bias from the NCO W_m. In addition, we assume that

$$\hat{\beta}_{3,m} \mid \beta_{3,m} \sim \mathcal{N}(\beta_{3,m}, \hat{\sigma}_m^2).$$

The result $\hat{\beta}_{3,m}$ and $\hat{\sigma}_m^2$ can be obtained directly from Equation (4). Applying this procedure across all NCOs and taking the average of the values of $\hat{\beta}_{3,m}$ provides an estimate $\hat{b}$, which serves as an estimate of the systematic bias b. This approach corresponds to the empirical posterior mean within the empirical Bayes framework. If $b = 0$, this suggests that the parallel trends assumption holds. Thus, we derive a test statistic for the parallel trends assumption and a corresponding two-sided test of the null hypothesis, $H_0 : b = 0$. The test statistic is derived from the distribution of estimated effects across multiple NCOs. The sign of b reveals the direction of bias in treatment effect estimates and the magnitude of estimated bias b indicates the severity of the violation. The p-value < 0.05 provides statistical evidence against parallel trends.

In the final step, we calibrate $\hat{\beta}_3$ by subtracting off the estimated bias, yielding the calibrated estimator:

$$\hat{\tau} = \exp(\hat{\beta}_3 - \hat{b}).$$

In the Supplement, we illustrate the complete details such as the statistical assumptions in Section C.5 and the theoretical guarantee of our proposed method in Section C.6.

Robust Estimation with Partially Invalid NCOs

When the minority of NCOs (less than 50%) are invalid—for example, when a selected NCO has some treatment effects, leading to additional bias in the estimation of τ—we adopt a more robust approach. Specifically, we use the median of $\hat{\beta}_{3,m}$ $(m = 1, \ldots, M)$ to compute $\hat{b}$. This adjustment ensures accuracy and robustness even in scenarios where the NCOs are not entirely reliable.

Data Generating Mechanism for Synthetic Data Experiments

We use $Y_{i,t}(z)$ to denote the potential outcome for the $Z = z$ $(z = 0, 1)$ group at the $T = t$ $(t = 0, 1)$ for the ith participant $(i = 1, \ldots, n)$. We assume the dimension of X_i is p, where the first p_1 elements are continuous, following a standard normal distribution. The rest of the elements of X_i are categorical, following a multinomial distribution. The dimension of the unmeasured confounders U is q. Similarly, we allow the types of unmeasured confounders to be a mix of continuous and categorical types.

The general data generation mechanism of the potential outcomes is displayed as:

$$\log\left[\mathbb{E}\{Y_{i,t}(0) \mid A_i, X_i, U_i\}\right] = \beta(A_i, X_i) + \gamma_t(U_i), \tag{4.5}$$

$$\log\left[\mathbb{E}\{Y_{i,1}(1) \mid X_i, U_i\}\right] = \beta(A_i, X_i) + \tau(X_i) + \gamma_1(U_i), \tag{4.6}$$

where $\beta(A_i, X_i)$, $\gamma_t(U_i)$, and $\tau(X_i)$ are individual-specific functions on X_i. $\beta(A_i, X_i)$ is a function specifying the influence of measured confounders on the potential outcomes for the intervention and control groups. $\gamma_t(U_i)$ captures the time-varying effect of unmeasured confounders on the potential outcomes. The $\tau(X_i)$ function specifies the treatment effect given time $T = 1$.

We conducted the simulation in two parts. The first part is about the type-I error rate and power for hypothesis testing for the parallel trends assumption. We ranged the sample sizes at 1000, 2000, and 3000. Similarly, for each sample size, the power is calculated for $b = 0.05$, 0.10, and 0.15.

The second part evaluates the bias and coverage probabilities for the proposed calibration method. We fix $\mathbb{E}[\gamma_1(U_i) - \gamma_t(U_i)] = 0.5$ and set the true value of the treatment effect, $\mathbb{E}[Y_1(1) - Y_1(0)]$, to be -1. We conducted method comparisons by calculating the relative bias with respect to the true

value. In this scenario, where the effect size is one, the relative bias is mathematically equivalent to the true bias, providing a direct assessment of estimation accuracy across all compared approaches.

The values for $\beta(A_i, X_i)$ and $\gamma_t(U_i)$ used in the first and second parts are shown in the Appendix (Section C.4). Additionally, for scenarios where a small proportion of invalid NCOs (10%) is included, we provide details on the generation of these invalid NCOs in Section C.7.

The replication number of the numerical experiments is set to 1000.

CHAPTER 5

RACIAL/ETHNIC DIFFERENCES IN LONG-COVID-ASSOCIATED SYMPTOMS AMONG PEDIATRICS POPULATION: FINDINGS FROM DIFFERENCE-IN-DIFFERENCES ANALYSES IN RECOVER PROGRAM

5.1. Introduction

The post-acute sequelae SARS-CoV-2 infection (PASC) has emerged as a significant concern (Davis et al., 2023, 2021; Thaweethai et al., 2023), particularly among young individuals with a previous diagnosis of COVID-19 (Organization et al., 2023; Morello et al., 2023; Selvakumar et al., 2023; Rodriguez-Morales et al., 2023; Zhou et al., 2024; Shen et al., 2025; Zhang et al., 2025a). Defined by the World Health Organization (WHO) as the persistence of at least one physical symptom for 12 weeks following initial testing without an alternative diagnosis and expanded by the National Institutes of Health (NIH) to include ongoing, relapsing, or new symptoms four or more weeks post-acute infection, PASC potentially affects a significant proportion of COVID-19 survivors (Lopez-Leon et al., 2021; Groff et al., 2021; Nasserie et al., 2021). In pediatric populations, recent studies show that PASC symptoms and conditions tend to be systematic and/or syndromic, with higher incidence conditions such as abnominal pain, loss of taste/smell, myocarditis, and symptoms associated with cold-like illness occurring in patients after the acute phase of COVID-19 (Rao et al., 2022; Wu et al., 2025; Zhang et al., 2025b,c; Li et al., 2025; Lu et al., 2025; Chen et al., 2025). The estimated prevalence of PASC symptoms and conditions in the pediatric population ranges from 23% to 45% among those previously infected by SARS-CoV-2 (Twohig et al., 2023; Lopez-Leon et al., 2022; O'Mahoney et al., 2023), depending on study designs and PASC definitions. These findings highlight the urgent need for further research and comprehensive support to address the prevalence of PASC in pediatric populations.

Prior investigations into potential racial/ethnic differences in PASC among adults have unearthed important findings. For example, the Centers for Disease Control and Prevention (CDC) have shown variations in PASC's impact based on race/ethnicity (National Center for Health Statistics et al.,

2022). A study by Khullar et al.16 reported that Non-Hispanic Black (NHB) individuals exhibited a higher incidence of new PASC symptoms and conditions compared to Non-Hispanic White (NHW) patients, a difference more pronounced in hospitalized than in non-hospitalized patients. These findings suggest the existence of potential racial/ethnic differences in PASC among adults. Importantly, it is crucial to note that race and ethnicity are social constructs rather than biological ones (Ford and Kelly, 2005; Ford and Harawa, 2010). Concurrently, research indicates children's likelihood of testing positive for COVID-19 correlates with their race/ethnicity (Saatci et al., 2021; Bandi et al., 2020; Swann et al., 2020; Lee et al., 2020). NHB, Hispanic, and multi-racial children exhibited higher rates of COVID-19 positivity compared to their NHW counterparts, indicating differences in infection rates across different racial/ethnic groups (Chisolm et al., 2022). However, limited research to date has addressed potential racial/ethnic differences in PASC among children and adolescents, making it a pressing area of study. Therefore, our study aimed to quantify such racial/ethnic differences by conducting an association study involving children and adolescents, to determine if the observed patterns are consistent with the findings from studies conducted among adults.

Examining health outcomes through the lens of racial/ethnic differences in the context of COVID-19 runs the risk of pre-existing racial/ethnic differences being either overshadowed or underestimated. To address this, we employed a difference-in-differences (DiD) approach to disentangle the shifts in racial/ethnic differences before and after COVID-19. Our investigation centered on a pediatric cohort from the RECOVER electronic health records (EHR) database across thirteen institutions. This study focused on the pediatric population with six months of follow-up, aims to investigate racial and ethnic differences in PASC symptoms and conditions attributable to SARS-CoV-2 infection. The study also included the Asian American/Pacific Islanders (AAPI) population, which was absent in the prior research.

Our main aim is to ascertain whether COVID-19 status correlates with any racial/ethnic differences across PASC symptoms and conditions. We use the term PASC symptoms and conditions to include a broad spectrum of health issues observed following COVID-19 infection (Rao et al., 2022;

Razzaghi et al., 2024), which are not exclusive to the post-COVID period. This includes systemic manifestations such as fatigue and malaise, respiratory symptoms, cardiovascular complications, and neurological disorders. Our definition, broader than the clinical diagnosis of PASC (ICD-10 code U09.9), enables us to investigate prevalence changes across different racial/ethnic groups before and after COVID-19 infection. Importantly, we quantified the racial/ethnic differences linked with COVID-19 by carefully accounting for pre-infection differences in PASC symptoms and conditions through the application of DiD analyses.

Specifically, Figure 5.1a demonstrates how to use the DiD model to find the racial/ethnic differences attributable to COVID-19. We measured the pre-infection racial/ethnic differences by comparing PASC symptoms and conditions between minority racial/ethnic groups and the NHW group during the pre-infection period. We then measured post-infection differences using the same metrics. The racial/ethnic differences attributable to COVID-19 were determined by calculating the difference between these post-infection and pre-infection differences. When COVID-19 may modify the effect of unmeasured confounding variables, we applied a negative control method (Figure 5.1c) to adjust for the shifted impacts of these potential unmeasured confounders. By addressing these questions, our objective is to shed light on racial/ethnic differences in the relationship between COVID-19 and PASC symptoms and conditions.

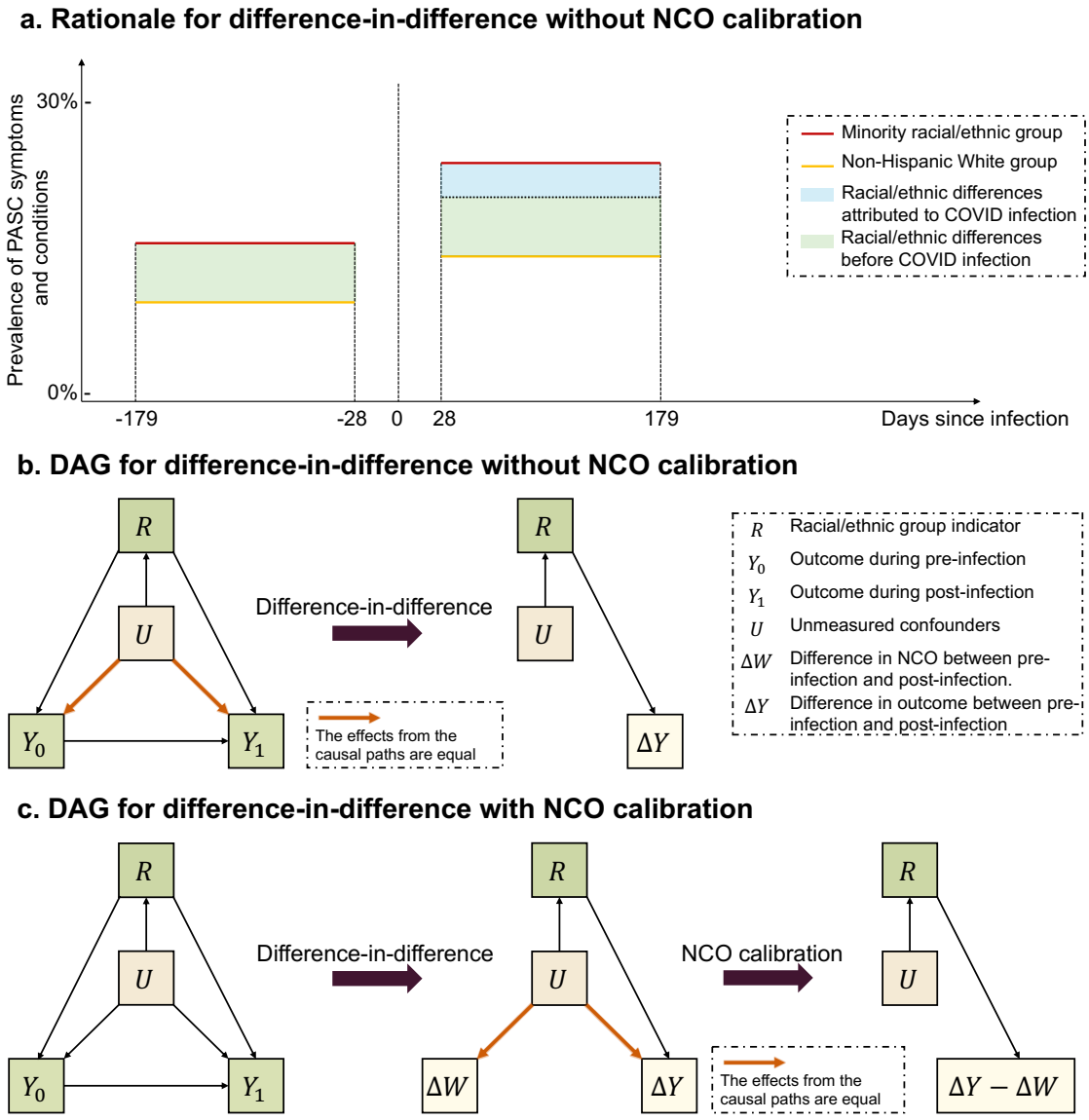

Figure 5.1: Illustration of difference-in-differences (DiD) analysis. (a) The rationale for DiD without the negative control outcome (NCO) calibration. It disentangles the racial/ethnic differences related to COVID-19 infection in PASC symptoms and conditions from the pre-infection observed racial/ethnic differences. (b) The directed acyclic graph (DAG) for the DiD without NCO calibration. Each node in the DAG represents each variable and the arrow symbol shows the causal effect. The left panel illustrates the parallel trends assumption for DiD. The right panel demonstrates how DiD without NCO calibration effectively blocks pathways from unmeasured confounders to PASC symptoms and conditions, provided the parallel trends assumption holds. (c) The DAG for the DiD with NCO calibration. The left panel illustrates a scenario where the parallel trends assumption is violated. The middle panel demonstrates that DiD without NCO calibration fails to block the pathway from unmeasured confounders to PASC symptoms and conditions. The right panel highlights that the DiD method with NCO calibration successfully eliminates the bias from unmeasured confounding.

5.2. Results

Study cohort and characteristics

The study involved 225,723 children and adolescents with COVID-19 (COVID-19-positive cohort) from March 2020 to October 2022 (cohort entry month). The index date for the cohort was the earliest date of documented SARS-CoV-2 infection, determined by a positive polymerase chain reaction (PCR), serology, antigen test, or diagnosis of COVID-19 or PASC (U09.9). The COVID-19-positive cohort comprised 109,022 NHW patients (48.3%), 45,823 NHB patients (20.3%), 60,012 Hispanic patients (26.6%), and 10,866 AAPI patients (4.8%) with SARS-CoV-2 infection from March 2020 to October 2022 (Figure 5.2). Of these, 50.2% were female.

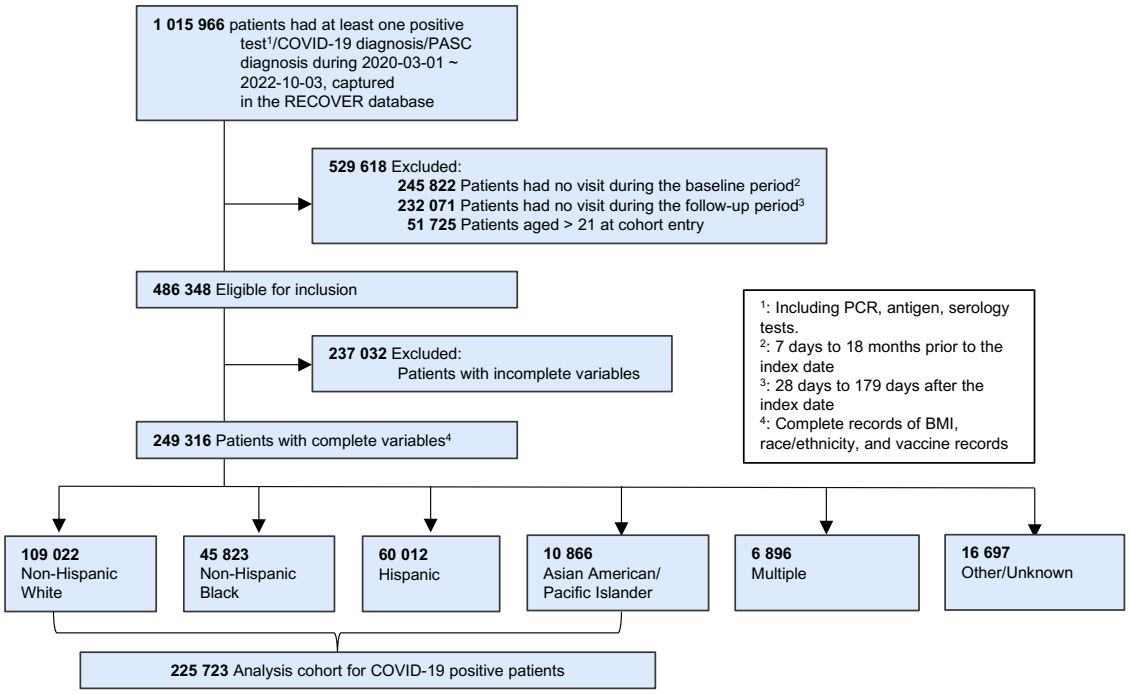

Figure 5.2: Flowchart for COVID-19 positive cohort. Flowchart illustrating the selection process for COVID-19-positive patients from March 2020 to October 2022. Abbreviation: BMI, body mass index; PASC, Post-Acute Sequelae of SARS-CoV-2 infection; PCR, Polymerase Chain Reaction.

We stratified the COVID-19-positive patients at cohort entry into two groups based on the severity during the acute phase (seven days before to fourteen days after the index date) of the COVID-19 infection (Forrest et al., 2022): the severe group (moderate and severe cases) and the non-severe

group (asymptomatic and mild cases). Table 5.1 and Table D.4 present the baseline characteristics of the cohort, such as age, sex, and Pediatric Medical Complexity Algorithm (Simon et al., 2018)(PMCA), stratified by race/ethnicity and severity. The definitions for each variable are shown in Table D.2.

Table 5.1: Baseline characteristics of COVID-19 positive patients by race/ethnicity and severity status

	Severe[1]				**Non-severe**[2]			
	NHW (N=9,140)	AAPI (N=590)	Hispanic (N=3,786)	NHB (N=3,346)	NHW (N=99,882)	AAPI (N=10,276)	Hispanic (N=56,226)	NHB (N=42,477)
Age categories								
<1 year	1601 (17.5%)	122 (20.7%)	767 (20.3%)	733 (21.9%)	9655 (9.7%)	965 (9.4%)	6154 (10.9%)	4686 (11.0%)
1 to <5 years	2811 (30.8%)	197 (33.4%)	1174 (31.0%)	910 (27.2%)	22018 (22.0%)	2802 (27.3%)	11892 (21.2%)	9147 (21.5%)
5 to <12 years	1893 (20.7%)	133 (22.5%)	773 (20.4%)	568 (17.0%)	27547 (27.6%)	3191 (31.1%)	17205 (30.6%)	12991 (30.6%)
12 to <16 years	1243 (13.6%)	60 (10.2%)	489 (12.9%)	482 (14.4%)	18276 (18.3%)	1433 (13.9%)	10430 (18.6%)	7721 (18.2%)
16 to <21 years	1592 (17.4%)	78 (13.2%)	583 (15.4%)	653 (19.5%)	22386 (22.4%)	1885 (18.3%)	10545 (18.8%)	7932 (18.7%)
Sex								
Female	4208 (46.0%)	255 (43.2%)	1728 (45.6%)	1580 (47.2%)	50682 (50.7%)	4968 (48.3%)	28412 (50.5%)	21408 (50.4%)
Male	4932 (54.0%)	335 (56.8%)	2058 (54.4%)	1766 (52.8%)	49200 (49.3%)	5308 (51.7%)	27814 (49.5%)	21069 (49.6%)
Obesity								
No	6055 (66.2%)	422 (71.5%)	2230 (58.9%)	1926 (57.6%)	62750 (62.8%)	6141 (59.8%)	21525 (38.3%)	21711 (51.1%)
Yes	3085 (33.8%)	168 (28.5%)	1556 (41.1%)	1420 (42.4%)	37132 (37.2%)	4135 (40.2%)	34701 (61.7%)	20766 (48.9%)
PMCA								
No condition	4333 (47.4%)	311 (52.7%)	1938 (51.2%)	1552 (46.4%)	67424 (67.5%)	7594 (73.9%)	41294 (73.4%)	27868 (65.6%)
Non-complex chronic	1435 (15.7%)	79 (13.4%)	534 (14.1%)	524 (15.7%)	18614 (18.6%)	1564 (15.2%)	8820 (15.7%)	8692 (20.5%)
Complex chronic	3372 (36.9%)	200 (33.9%)	1314 (34.7%)	1270 (38.0%)	13844 (13.9%)	1118 (10.9%)	6112 (10.9%)	5917 (13.9%)
Negative tests prior entry								
0	6038 (66.1%)	407 (69.0%)	2671 (70.5%)	2301 (68.8%)	73195 (73.3%)	7908 (77.0%)	45852 (81.5%)	32592 (76.7%)
1	1445 (15.8%)	97 (16.4%)	521 (13.8%)	497 (14.9%)	15999 (16.0%)	1440 (14.0%)	6650 (11.8%)	6233 (14.7%)
2	641 (7.0%)	32 (5.4%)	225 (5.9%)	238 (7.1%)	5875 (5.9%)	519 (5.1%)	2169 (3.9%)	2146 (5.1%)
≥3	1016 (11.1%)	54 (9.2%)	369 (9.7%)	310 (9.3%)	4813 (4.8%)	409 (4.0%)	1555 (2.8%)	1506 (3.5%)
Vaccine dosage								
0	7860 (86.0%)	469 (79.5%)	3347 (88.4%)	3015 (90.1%)	82136 (82.2%)	7256 (70.6%)	46760 (83.2%)	37252 (87.7%)
1	240 (2.6%)	22 (3.7%)	80 (2.1%)	71 (2.1%)	2878 (2.9%)	474 (4.6%)	1667 (3.0%)	1159 (2.7%)
≥2	1040 (11.4%)	99 (16.8%)	359 (9.5%)	260 (7.8%)	14868 (14.9%)	2546 (24.8%)	7799 (13.9%)	4066 (9.6%)

Abbreviations: NHW: Non-Hispanic White; NHB: Non-Hispanic Black; AAPI: Asian American or Pacific Islander; PMCA: Pediatric Medical Complexity Algorithm.
[1]Moderate/severe COVID-19; [2]Asymptomatic/mild COVID-19.

Statistical method overview

The PASC symptoms and conditions were defined using ICD-10-CM codes (shown in Table D.2) and were measured during the pre-infection period (28 to 179 days before the index date) and the post-infection period (28 to 179 days after the index date). We performed pairwise propensity score matching for each minority racial/ethnic group (NHB, Hispanic, AAPI) to the NHW group based on pre-specified covariates. To adjust for pre-infection racial/ethnic differences, we applied a DiD model with negative control outcome (NCO) calibration (Figure 5.1c) to the matched cohort.

Figure 5.1 illustrates the rationale and methodology of our difference-in-differences (DiD) with the

NCO calibration approach. Figure 5.1b shows the rationale of the DiD model. The parallel trends assumption in the DiD model within our study states that the racial/ethnic differences do not change over time in the absence of COVID-19 infection. When the parallel trends assumption holds, unmeasured confounding variables do not affect the racial/ethnic difference in outcomes before and after COVID-19 infection. This is represented by the absence of an arrow between the unmeasured confounding variables and the differences in outcomes in Figure 5.1b. However, when the parallel trends assumption is violated—specifically if COVID-19 modifies the effect of unmeasured confounding variables—the outcome estimates from the DiD may become biased without the NCO calibration. To address this limitation, we implemented negative control experiments for each minority racial/ethnic group compared to the NHW group, as illustrated in Figure 5.1c. This approach allows us to calibrate our estimates from the unmeasured confounding bias. We provide details for explanations for Figure 5.1b and Figure 5.1c in the Methods Section.

Racial/ethnic differences in PASC symptoms and conditions

To investigate racial/ethnic differences attributable to COVID-19, we examined twenty-four PASC symptoms and conditions. We grouped the PASC symptoms and conditions into two categories: systemic and syndromic features (shown in Table D.3). Syndromic features encompass symptoms, signs, and non-specific laboratory abnormalities, while systemic features are diagnosed health conditions. We used thirty-one NCOs in this study, defined as outcomes assumed to have no racial/ethnic differences attributable to COVID-19 (Table D.1).

After achieving the balance of standardized mean difference (SMD) by propensity score matching and conducting the DiD model with NCO calibration, Figure 5.3 shows the racial/ethnic differences attributable to COVID-19 in PASC symptoms and conditions stratified by the severity of acute COVID-19 after the NCO calibration. Overall, we found evidence of an increase in composite outcomes, i.e., at least one condition and any of the syndromic conditions, after SARS-CoV-2 infection among the AAPI group in both severe and non-severe COVID-19 group, but there was no evidence of increased racial differences among Hispanic and NHB groups.

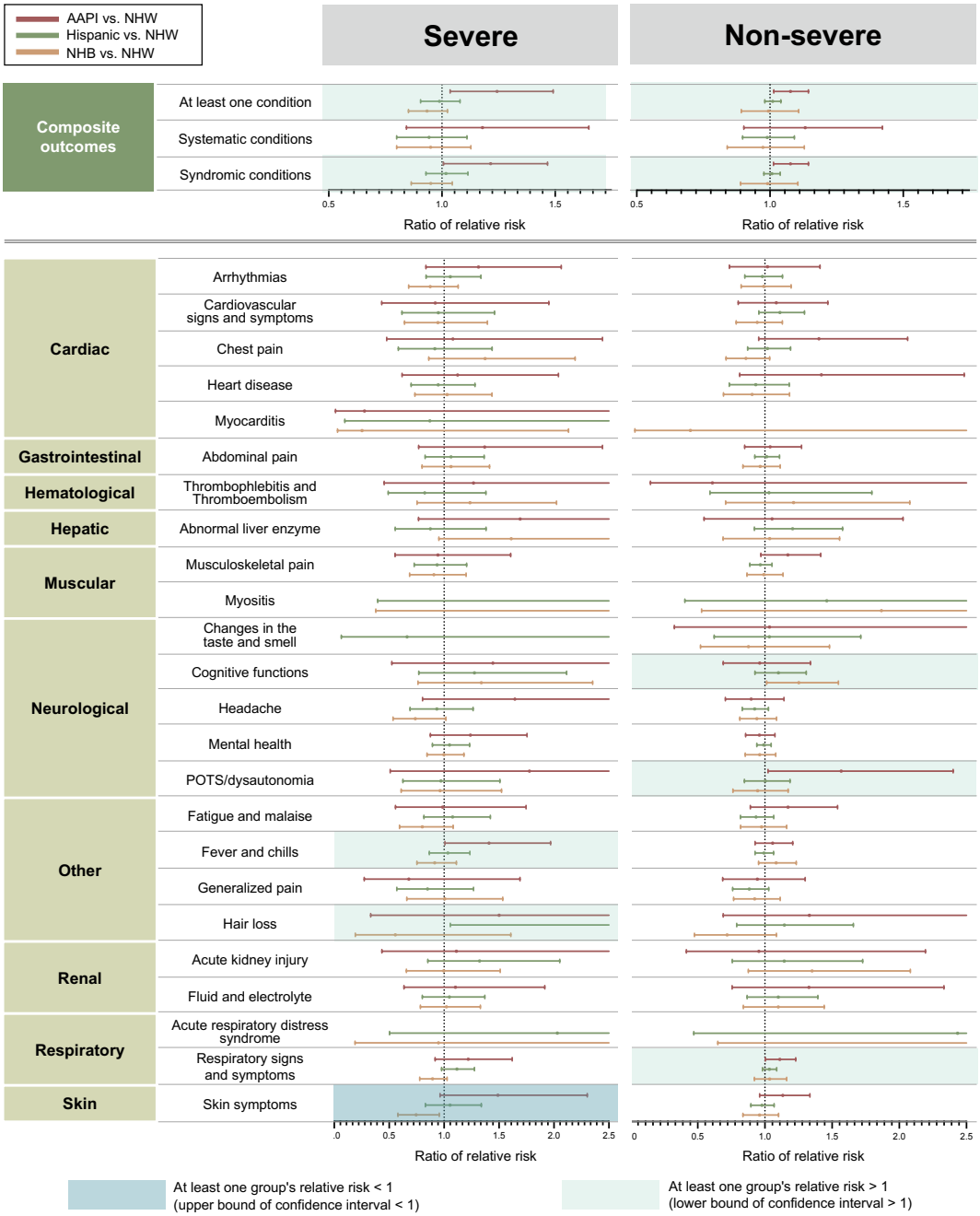

Figure 5.3: Racial/Ethnic differences attributable to COVID-19 in PASC symptoms and conditions Racial/Ethnic differences in risk ratio (RR) that are attributable to COVID-19 for the prevalence of PASC symptoms and conditions among COVID-19-positive cohort (n=225,723 participants), by race/ethnicity and severity status. Each minority group—Asian American/Pacific Islanders (AAPI), Non-Hispanic Black (NHB), and Hispanic—was compared to the Non-Hispanic White (NHW) group. The dots in the figure showed the estimated RRs for each minority group versus the NHW group from the difference-in-differences (DiD) analysis with NCO calibration stratified by severity status. The error bars showed the 95% confidence interval (CI) of the estimated RR.

For patients with severe COVID-19, AAPI patients showed a higher increase in at least one condition (risk ratio [RR] 1.24, 95% confidence interval (CI) 1.04 to 1.49, two-sided Wald test P=0.019) and any of syndromic conditions (RR 1.22, 95% CI 1.01 to 1.47, P=0.042) compared to NHW after SARS-CoV-2 infection. Both Hispanic patients (RR 0.99, 95% CI 0.91 to 1.08, P=0.804) and NHB patients (RR 0.93, 95% CI 0.85 to 1.24, P=0.147) showed a non-significant decrease in at least one condition as compared to NHW patients. For patients with non-severe COVID-19, AAPI patients showed an increase in at least one condition (RR 1.08, 95% CI 1.01 to 1.14, P=0.015) and any of syndromic conditions (RR 1.08, 95% CI 1.01 to 1.08, P=0.017) compared to NHW. Hispanic patients showed a non-significant increase in at least one condition (RR 1.01, 95% CI 0.98 to 1.04, P=0.498), while NHB patients had a similar non-significant risk (RR 0.99, 95% CI: 0.89 to 1.11, P=0.915) in at least one condition.

However, there exist significant differences among all minority groups across several PASC symptoms and conditions after SARS-CoV-2 infection. For example, for patients with severe COVID-19, the increased prevalence of hair loss among Hispanic patients was greater (RR 2.62, 95% CI 1.06 to 6.49, two-sided Wald test P=0.038) than the increased prevalence among NHW patients. The corresponding increase in the prevalence of fever and chills among AAPI was greater (RR 1.41, 95% CI 1.01 to 1.97, P=0.045) than NHW. NHB patients had a smaller change in skin symptoms (RR 0.74, 95% CI 0.58 to 0.96, P=0.021) than NHW patients. For patients with non-severe COVID-19, AAPI patients had a greater change concerning POTS/dysautonomia (RR 1.57, 95% CI 1.02 to 2.40, P=0.037) and respiratory signs and symptoms (RR 1.11, P=0.036) with 95% CI 1.00 to 1.23 (significant; rounded to two decimals) compared to NHW patients. NHB patients had an increase in cognitive function (1.25, 95% CI 1.01 to 1.55, P=0.037) than NHW patients.

Furthermore, we observed that these racial/ethnic differences varied depending upon the severity of the acute phase of COVID-19 as well as the specific PASC symptoms and conditions being analyzed. For example, among the severe group, the differential increase in abdominal pain was more pronounced for all three minority groups compared to those in the non-severe category.

Sensitivity analysis

Figure D.3 showed the racial/ethnic differences using standard Poisson regression models. Among COVID-19 patients within the severe group, NHB patients showed a greater incidence in at least one condition (RR 1.16, 95% CI 1.02 to 1.32, two-sided Wald test P=0.024) and any syndromic conditions (RR 1.14, P=0.042) with 95% CI 1.00 to 1.30 (significant; rounded to two decimals) as compared to NHW patients. Hispanic patients also showed a non-significant greater incidence in at least one condition (RR 1.12, 95% CI 0.99 to 1.27, P=0.075) as compared to NHW patients.

In Figure D.3, among COVID-19 patients with severe illness during the acute infection, Hispanic individuals exhibited a greater incidence of respiratory signs and symptoms (RR: 1.16, 95% CI 1.02 to 1.33, two-sided Wald test P=0.024) and hair loss (RR: 1.84, 95% CI 1.02 to 3.31, P=0.043) as compared with the NHW patient group. NHB had a greater incidence of respiratory signs and symptoms (RR: 1.19, 95% CI 1.03 to 1.36, P=0.015) and heart disease (RR: 1.48, 95% CI 1.06 to 2.07, P=0.022), but a lower incidence of arrhythmias (RR: 0.73, 95% CI 0.57 to 0.94, P=0.013) and headache (RR: 0.66, 95% CI 0.48 to 0.93, P=0.016) compared with the NHW group. Among those with non-severe acute COVID-19, Hispanic patients displayed a higher incidence of myocarditis (RR 4.28, 95% CI 1.53 to 11.98, P=0.006) and abnormal liver enzyme (RR 2.06, 95% CI 1.08 to 3.94, P=0.029) compared with NHW patients. Meanwhile, AAPI patients demonstrated a greater incidence of hair loss (RR 3.32, 95% CI 1.43 to 7.72, P=0.005) compared with the NHW patient group.

These findings revealed that our DiD approach identified fewer racial/ethnic differences compared to standard regression models. It is worth noting that the DiD approach adjusted for the baseline racial/ethnic differences before the SARS-CoV-2 infection, a step that a standard regression analysis failed to take into consideration. Consequently, some of the observed racial/ethnic differences with prior work might not be attributed to COVID-19. Nevertheless, given its adjustment for baseline racial/ethnic differences, the DiD approach holds greater robustness.

To determine whether the defined set of twenty-four PASC symptoms and conditions accurately reflect PASC, we conducted a crude incidence analysis comparing these symptoms and conditions in the COVID-19-positive cohort to that in the COVID-19-negative cohort. A total of 677,448

COVID-19-negative patients were included in the crude analysis. A random negative test date was chosen as the index date for COVID-19-negative patients and the selection criteria for the COVID-19-negative cohort are shown in Figure D.1. The results indicated that the incidence of all PASC symptoms and conditions was higher in all COVID-19-positive patients compared to all COVID-19-negative patients (Table D.6).

To further validate our definition of PASC symptoms and conditions, as well as to investigate potential racial/ethnic differences in PASC symptoms attributable to COVID-19, we applied the DiD approach with NCO calibration to the COVID-19-negative cohort (Figure D.2). This analysis revealed no racial or ethnic differences in PASC symptoms and conditions attributable to COVID-19 within the COVID-19-negative cohort. These findings suggest that the observed racial/ethnic differences in PASC symptoms and conditions among COVID-19-positive individuals are unlikely to be due to inherent differences in symptom and condition incidence patterns but rather may reflect factors specifically associated with COVID-19 infection or its sequelae. This analysis validates that the racial and ethnic differences in PASC observed in the COVID-19-positive cohort are related to COVID-19 infection, rather than underlying population differences or unrelated health trends.

5.3. Discussion

We examined racial/ethnic variations in long-term consequences of documented SARS-CoV-2 infection across thirteen health institutions in the RECOVER study for 225,723 children and adolescents. After accounting for pre-existing differences and potential confounding biases, we found evidence suggesting differences attributable to COVID-19 for AAPI patients compared to NHW patients. Notably, AAPI patients showed an increase in the risk of developing at least one PASC-related condition and any syndromic conditions, both in severe and non-severe COVID-19 cases. We observed specific symptom differences among minority racial/ethnic groups, despite no evidence indicating differences in composite outcomes for NHB or Hispanic populations compared to NHW. For example, NHB patients showed a smaller change in skin symptoms compared to NHW patients, consistent with previous findings in adults by Khullar et al. (2023).

The increased cognitive function issues in NHB patients, the greater prevalence of hair loss in

Hispanic patients, the higher risk of POTS or dysautonomia, and respiratory symptoms in AAPI patients with severe COVID-19 all warrant particular attention in clinical settings. It is worth noting that AAPI is the smallest racial/ethnic group, and for this group, the results show evidence of differences compared to NHW. Therefore, this strengthens the findings of racial/ethnic differences across PASC symptoms and conditions for the other minority groups compared to NHW.

While the observed racial/ethnic differences in RRs are subtle, they remain significant when applied to large populations and may reflect broader systemic issues. For example, the observed differences in PASC can primarily be attributed to differential healthcare access and the availability of specialized pediatric care, which vary significantly across different regions of the country, according to a national report29. These variations are not due to biological reasons but are more closely linked to socioeconomic factors, healthcare infrastructure, and regional healthcare policies.

This study examines PASC racial/ethnic differences among pediatric populations, providing valuable insights into the clinical relevance of these differences. Understanding these differences is crucial for developing targeted interventions to ensure equitable healthcare access and outcomes for all pediatric populations affected by PASC. The increased risk among minority racial/ethnic pediatric patients may necessitate targeted follow-up care and support for the minority racial/ethnic population. Healthcare providers should be aware of the potential for varied PASC presentations across racial/ethnic groups in pediatric populations. This knowledge can inform more targeted screening and follow-up protocols, potentially improving early detection and management of PASC symptoms in different racial/ethnic patient populations.

The implications of this study on a broader public health level are significant. By utilizing data from thirteen institutions, representing approximately 10% of the US children population and encompassing both urban and suburban health systems, this study offers a comprehensive national sample. Public health strategies are suggested to prioritize the reduction of healthcare access and the enhancement of specialized pediatric care availability across all regions. This will help in better managing and treating PASC in children, ensuring that all affected patients receive the necessary care regardless of their geographic location or socioeconomic status. Furthermore, these findings

underscore the importance of continuous surveillance, research, and tailored public health policies to address the evolving needs of this patient group.

The method utilized in our study has multiple strengths. First, we used propensity score matching methods instead of standard linear regression models in our adjustment of the confounders, which helped us reduce the non-linear effects of the confounders (Amoah et al., 2020). Second, we accounted for the pre-infection racial/ethnic differences in PASC symptoms and conditions. This approach enabled us to quantify racial/ethnic differences attributable to COVID-19 more accurately. The DiD model is particularly powerful in the context of studying racial/ethnic differences, as it helps to isolate the effect of COVID-19 from pre-existing ones. By comparing the change in health outcomes before and after COVID-19 across racial/ethnic groups, we can more confidently attribute observed differences to the impact of the virus rather than to the pre-infection racial/ethnic differences.

A key strength of our study is the implementation of NCOs, which are outcomes not expected to show racial/ethnic differences due to COVID-19. This approach enhances the reliability and interpretability of our findings. By analyzing these alongside our primary outcomes, we enable the detection of systematic bias from unmeasured confounding variables. This approach allows us to calibrate results from the DiD model. Importantly, NCOs help mitigate the impact of unmeasured confounding that may change before and after the infection. This is particularly crucial when studying racial/ethnic differences, where many dynamic social and environmental factors may not be captured in our dataset. Furthermore, we carefully addressed potential bias related to differences in follow-up times. To ensure comparability, we standardized the follow-up period for all participants, observing PASC symptoms and conditions over a consistent timeframe of five months (28 to 179 days post-cohort entry). This approach mitigates any bias that might arise from variations in follow-up duration across individuals.

This study has several limitations. First, socioeconomic differences may exacerbate racial/ethnic differences in PASC symptoms and conditions, thereby acting as a mediator effect in the causal pathway between race/ethnicity and clinical outcomes. Such influences have been suggested as risk

factors for acute COVID-19 by Chisolm and colleagues in a RECOVER EHR study (Chisolm et al., 2022). Future research on PASC outcomes is of interest to study such mediation effects.

Secondly, health-seeking behavior or healthcare access is an important consideration, known as ascertainment bias (Asch et al., 2021b). Limited access to care and associated medical records among certain minority racial/ethnic groups may contribute to potential bias in the observed racial/ethnic differences. Related issues were recently described by Nasir et al. (2023) for the ascertainment of PASC symptoms and conditions in adult populations through EHR. Nevertheless, we tried to address the ascertainment bias by including healthcare utilization factors in the propensity score model, such as the number of inpatient, outpatient, and ED visits during the baseline period.

Thirdly, confounding poses a significant bias threat in observational studies. To address this, we extensively adjusted for potential confounders using a propensity-score-based matching method and DiD analyses. We employed NCOs to reduce the residual bias, such as unmeasured confounder bias. Additionally, EHR data completeness issues may lead to misclassification and loss-to-follow-up bias. Some attempts have been made to mitigate the impacts of these biases (Duan et al., 2017; Chen et al., 2019; Schuemie et al., 2014, 2018). The analysis used a combined set of patients, but potential bias in racial/ethnic differences may vary between outpatients and inpatients. Addressing these issues can help improve the reliability of evidence generated from these investigations.

Fourthly, asymptomatic cases are less likely to be captured in EHR as they are less likely to be tested, impacting the findings for non-severe COVID-19. This study focused on diagnosed or treated cases of PASC, reflecting children who sought healthcare services. The direction of bias can be bidirectional, depending on whether asymptomatic cases are disproportionately missed in minority or NHW groups. This could attenuate the DiD analysis for the non-severe COVID-19 group due to potential misclassification bias, leading to conservative or inflated results (Wu et al., 2024b). Future research should aim to capture a more comprehensive spectrum of COVID-19 severity.

Fifthly, our study design, requiring at least one healthcare visit within 28-179 days after the index date, balanced selection bias against loss to follow-up bias (Howe et al., 2016). While this approach

may overrepresent frequent healthcare users, especially during 2020-2021, we mitigated this by incorporating diverse testing data to reduce COVID-19 status misclassification. Furthermore, in the crude incidence analysis comparing COVID-19 positive and negative patients, the resulting selection bias likely included sicker patients in the COVID-19 negative group, potentially biasing our results towards the null. Nevertheless, we still observed a higher incidence of PASC symptoms and conditions for COVID-19-positive patients compared to COVID-19-negative patients. Future studies could explore alternative designs or data sources to enhance the external validity of these findings.

In summary, we quantified the racial/ethnic differences in PASC symptoms and conditions and the impact of SARS-CoV-2 infection on these differences. The impact of COVID-19 varied across racial/ethnic groups, severity of acute COVID-19, and different PASC symptoms and conditions.

5.4. Methods

Ethics and inclusion

This study constitutes human subject research. Institute Review Board (IRB) approval was obtained under Biomedical Research Alliance of New York (BRANY) protocol 21-08-508. As part of the BRANY IRB process, the protocol has been reviewed in accordance with the institutional guidelines. The BRANY waived the need for patient-informed consent and HIPAA authorization.

Data source

This retrospective cohort study is part of the NIH Researching COVID-19 to Enhance Recovery (RECOVER) Initiative (https://recovercovid.org/), which aims to learn about the long-term effects of COVID-19.

The data were contributed by 13 sites. Participating institutions in this study included: Children's Hospital of Philadelphia, Cincinnati Children's Hospital Medical Center, Children's Hospital of Colorado, Ann and Robert H. Lurie Children's Hospital of Chicago, Nationwide Children's Hospital, Nemours Children's Health System (in Delaware and Florida), Duke University, University of Iowa Healthcare, University of Michigan, University of Missouri, Oregon Community Health Informa-

tion Network (OCHIN), University of California, San Francisco, and Vanderbilt University Medical Center. All analyses were conducted using R version 4.1.2, with statistical significance set at a two-tailed p-value threshold of < 0.05.

Cohort construction

We conducted a retrospective study from March 1, 2020, to October 3, 2022, with at least 6 months of follow-up time. We included patients under the age of 21 who had at least one visit from 18 months before to 7 days before the index date (defined as the baseline period) and at least one encounter within 28 days and 179 days after the index date (defined as the follow-up period). For COVID-19-positive patients, we included the patients who had positive PCR, serology, or antigen tests or diagnoses of COVID-19, or diagnoses of PASC (U09.9), which we defined as documented SARS-CoV-2 infection. The index date for these patients was defined as the first time of SARS-CoV-2 infection. We used patient race/ethnicity data included in the EHR and categorized patients into six racial/ethnic groups: NHW, NHB, Hispanic, AAPI, Multiple, and Other/Unknown. Multiple comprising less than 3% of the population in any COVID-19 and severity category were considered small samples and excluded from the analysis. The Other/Unknown categories were also excluded due to interpretability issues. The selection of participants for COVID-19-positive patients in real-world data is summarized in Figure 5.2.

Defining outcomes

Our definition of PASC symptoms and conditions included 24 symptoms and conditions as shown in Rao et al.11, including abdominal pain, abnormal liver enzyme, acute kidney injury, acute respiratory distress syndrome, arrhythmias, cardiovascular signs and symptoms, changes in taste and smell, chest pain, cognitive functions, fatigue and malaise, fever and chills, fluid and electrolyte, generalized pain, hair loss, headache, heart disease, mental health disorders, musculoskeletal pain, myocarditis, myositis, Postural Orthostatic Tachycardia Syndrome (POTS) or dysautonomia, respiratory signs and symptoms, skin symptoms, and thrombophlebitis and thromboembolism. We used validated diagnostic codes (ICD-10-CM) confirmed by board-certified pediatricians, with details of the code sets available in the Supplementary Materials Table D.2. Systematic and syndromic condi-

tions related to PASC were grouped by the 24 PASC symptoms and conditions, which are detailed in Table D.4.

Patient characteristics

The primary exposure was race/ethnicity, categorized into NHW, NHB, Hispanic, and AAPI. Various patient characteristics were considered as confounders, including age at cohort entry date ($<$5, 5–12, 12–21), sex (female, male), cohort entry month (from March 2020 to October 2022), site indicators, obesity (obese, non-obese), a chronic condition indicator as defined by the PMCA (no chronic condition, non-complex chronic condition, complex chronic condition)(Simon et al., 2018), healthcare visits (inpatient, outpatient, emergency department visits), medications (0, 1, 2, $\geq$3), negative tests (0, 1, 2, $\geq$3), vaccine doses (0, 1, $\geq$2), and immunization duration during the baseline period (no vaccine, $<$4 months, $\geq$4 months). The severity of COVID-19 at cohort entry date was stratified into the following categories: asymptomatic, mild (symptomatic), moderate (involving moderately severe COVID-19-related conditions such as gastroenteritis, dehydration, and pneumonia), and severe (comprising unstable COVID-19-related conditions, ICU admission, or mechanical ventilation)(Forrest et al., 2022). Patients exhibiting asymptomatic or mild symptoms were categorized into the non-severe group, while all other patients were classified into the severe group.

Propensity score matching

To quantify the racial/ethnic differences in the PASC symptoms and conditions, we use RR as the comparative measure. The RR is known to be a collapsible measure, where collapsibility (Greenland et al., 1999) refers to the measure of association conditional on some factors that remain consistent with the marginal measure collapsed over strata (Whitcomb and Naimi, 2021).

To eliminate the impact of potential measured confounders, we used a propensity score matching technique with the covariates detailed in the patient characteristics section. The propensity score is calculated by the logistic regression model fitted by regressing the racial/ethnic groups on the covariates. We performed this matching separately for minority racial/ethnic groups (NHB, Hispanic, and AAPI), each stratified by severity status, compared with the NHW group. After performing the matching, we assessed the SMD between each covariate value for different racial/ethnic groups,

with a difference of 0.1 or less indicating an acceptable balance (Austin, 2009).

DiD models

Subsequently, we quantified the differential increase in the prevalence of PASC symptoms and conditions across different racial/ethnic groups using the DiD method on the matched cohort. Figure 5.1a provides a visual representation of the DiD method used to estimate racial/ethnic differences in the increased prevalence of PASC symptoms and conditions related to COVID-19. We introduce notation to illustrate the DiD method. For each minority racial/ethnic group (AAPI, NHB, or Hispanic) matched with the NHW cohort, we use a binary indicator R to denote racial/ethnic groups ($R = 0$ for NHW, $R = 1$ for NHB, Hispanic, or AAPI). We define T as the time period, where $T = 1$ represents the post-infection period (28 to 179 days after COVID-19 infection) and $T = 0$ represents the pre-infection period (28 to 179 days before COVID-19 infection). Y is a binary indicator denoting the occurrence of PASC symptoms and conditions, where $Y = 1$ indicates the patient has been diagnosed with specific PASC symptoms and conditions, and $Y = 0$ indicates the absence of such a diagnosis. For each participant, we observed the outcome Y in both pre-infection ($T = 0$) and post-infection ($T = 1$) periods.

We fitted a Poisson regression model by regressing each PASC symptom or condition on racial/ethnic groups, time period, and the interaction between racial/ethnic groups and time period:

$$\log\left(E[Y \mid R, T]\right) = \beta_0 + \beta_1 R + \beta_2 T + \beta_3 (R \times T), \tag{5.1}$$

where $\log(\cdot)$ is the log link function. From Equation (5.1), the pre-infection racial/ethnic difference is found by the coefficient β_1, as shown in the green part in Figure 5.1a. The post-infection racial/ethnic difference is denoted by the sum of coefficients $\beta_1 + \beta_3$. Thus, the coefficient of the interaction term, β_3, represents the differential increase in racial/ethnic differences attributable to COVID-19, shown in the blue part in Figure 5.1a.

To illustrate that the DiD model removes bias from the unmeasured confounder when the parallel trends assumption holds, Figure 5.1b shows the directed acyclic graph (DAG) for the causal re-

lationship (Pearl, 2009) when conducting the DiD model. We use Y_0 for the outcome during the pre-infection period, Y_1 for the outcome during the post-infection period, and U for the unmeasured confounding variables. For simplicity, we present the DAG for the matched cohort of each minority race/ethnicity-NHW pair after adjusting for measured confounding variables X using the matching method.

Arrows between variables denote causal relationships. For instance, the arrow from R to Y_1 represents the racial/ethnic difference during the post-infection period, while the arrow from R to Y_0 represents the difference during the pre-infection period. The DiD model aims to identify the racial/ethnic difference in the outcome change between the post-infection and pre-infection periods, ΔY (where $\Delta Y = Y_1 - Y_0$), as shown by the arrow from R to ΔY.

The key advantage of the DiD model is its ability to adjust for unmeasured confounding variables U under the parallel trends assumption. This assumption states that the effect of U remains constant over time, meaning the effect of U on Y_0 and Y_1 is equal. In the DAG, this is represented by the paths $U \rightarrow Y_0$ and $U \rightarrow Y_1$ being equal.

Under this assumption, the DiD model, by taking the difference between Y_1 and Y_0, effectively blocks the path from U to ΔY, allowing us to isolate the effect of R on ΔY without interference from U. Notably, a specified model adjusting for pre-infection outcomes Y_0 cannot adjust for the bias from the unmeasured confounder U. Such a model effectively blocks the path $R \rightarrow Y_0 \rightarrow Y_1$ but introduces new bias through the collider path $R \rightarrow Y_0 \leftarrow U \rightarrow Y_1$. Collider bias occurs when two variables, such as R and U, influence a common effect, such as Y_0, and controlling for this effect inadvertently creates a spurious association, leading to biased results. This is why the DiD method is more accurate than the model adjusting for the outcome during the pre-infection period.

Negative control outcome calibration

While we used propensity score matching to account for measured confounders and DiD analyses to address pre-infection racial/ethnic differences, results can still be impacted by unmeasured confounder bias when the parallel trends assumption does not hold. To mitigate such bias, we collected

thirty-one NCOs prespecified by pediatric physicians, not expected to show racial/ethnic differences due to COVID-19. We fitted Equation (5.1) with each NCO as the outcome. Residual bias thus can be estimated from the average of β_3s from NCO models. By using these NCOs, the study calibrated residual bias from unmeasured and systematic sources.

Figure 5.1c illustrates this using an expanded DAG. When the parallel trends assumption is violated, a backdoor path from U to ΔY can emerge, potentially biasing our estimates. To address this, we incorporate NCOs into our analysis, denoting each NCO during post- and pre-infection periods by W_1 and W_0, respectively. We assume the effect of U on ΔW (where $\Delta W = W_1 - W_0$) equals its effect on ΔY. By subtracting ΔW from ΔY, we effectively block the path from U to $\Delta Y - \Delta W$, enhancing accuracy and robustness against unmeasured confounding. We conducted a two-sided Wald test to evaluate racial/ethnic differences and reported corresponding p-values without adjustments for multiple comparisons.

Sensitivity analysis

We conducted a list of sensitivity analyses to examine the robustness of our findings. To evaluate how different statistical methods might influence the analytical results, we used an alternative approach, standard Poisson regression analyses, with RR as the comparative measure. Specifically, we considered the incidence of PASC symptoms and conditions as outcomes, while controlling for the same confounders that were used in the matching process within the DiD analysis. The p-values are reported from two-sided Wald tests. In addition, we performed crude incidence analyses to compare PASC symptoms and conditions between COVID-19 positive and negative groups, assessing whether the prespecified outcomes accurately reflect PASC. We showed the p-values for the statistical significance from two-sided two-proportion Z-tests. To validate racial/ethnic differences attributed to COVID-19, we conducted DiD analyses with NCO calibration on COVID-19 negative groups.

Code Availability

The code used for the analysis in this study is available and can be accessed in the Code Ocean repository at https://doi.org/10.24433/CO.0460366.v1.

CHAPTER 6

EMPOWERING TRIAL BY MULTI-SOURCE MULTI-SITE REAL-WORLD DATA: A NEGATIVE CONTROL-CALIBRATED DIGITAL TWIN APPROACH

6.1. Introduction

Randomized controlled trials (RCT) offer high internal validity through randomization, minimizing confounding and ensuring robust causal inference. However, their restrictive inclusion criteria and controlled settings often limit generalizability, and their high costs constrain scalability. In contrast, Real-world data (RWD) represent complementary sources of evidence in clinical research (Rothwell, 2005; Abernethy et al., 2017). RWD, encompassing electronic health records (EHR), insurance claims, and patient registries, provides extensive population coverage and longitudinal follow-up but is subject to selection bias and residual confounding due to the absence of randomization (Sherman et al., 2016; Abernethy et al., 2017). RCTs provide structured and complete access to predefined data attributes, whereas RWD often offers only partial access, particularly when derived from EHRs (Cowie et al., 2017). Moreover, while RCTs typically have moderate sample sizes and high costs, RWD enables large-scale analyses with lower financial constraints. Leveraging external controls derived from RWD can significantly enhance statistical efficiency for RCT, potentially reducing required sample sizes, shortening study durations, and lowering overall costs (Viele et al., 2014; Schmidli et al., 2020). Therefore, the strategic integration of RCTs with RWD represents a promising approach to achieve both methodological rigor and cost-effective, broadly applicable clinical research.

Digital twin technology has gained attention as a promising solution for enhancing causal inference in clinical research (Katsoulakis et al., 2024). Originally developed in engineering, digital twins leverage machine learning and high-dimensional RWD to create individualized virtual representations of patients, allowing researchers to model counterfactual outcomes and simulate treatment responses in real-world settings. This approach has been applied in biomedical research to predict disease progression, optimize treatment strategies, and personalize interventions (Venkatesh et al.,

2022). The PROCOVA method enhances RCTs by training digital twins from RWD to construct a virtual control group for each participant, thereby increasing statistical efficiency (Schuler et al., 2022).

Despite the promise of digital twins in strengthening RCTs, the validity of digital twin-based analyses depends on precise calibration, as discrepancies between trial and real-world populations can introduce systematic biases (Figure 6.1a). Existing methods, such as negative control outcomes and synthetic control techniques, offer potential strategies for mitigating these biases, yet challenges remain in ensuring accurate transportability of trial-based estimates to real-world populations.

To address these challenges, we propose a novel framework that systematically corrects for model shift bias in RCT-RWD integration using a digital twin approach with negative control outcomes. We introduce a data-driven calibration procedure that leverages negative control outcomes to quantify and adjust for systematic biases, ensuring that treatment effect estimates remain robust across different data sources. This approach offers a principled methodology for generating reliable real-world evidence, facilitating more accurate and generalizable insights from combined RCT and RWD analyses.

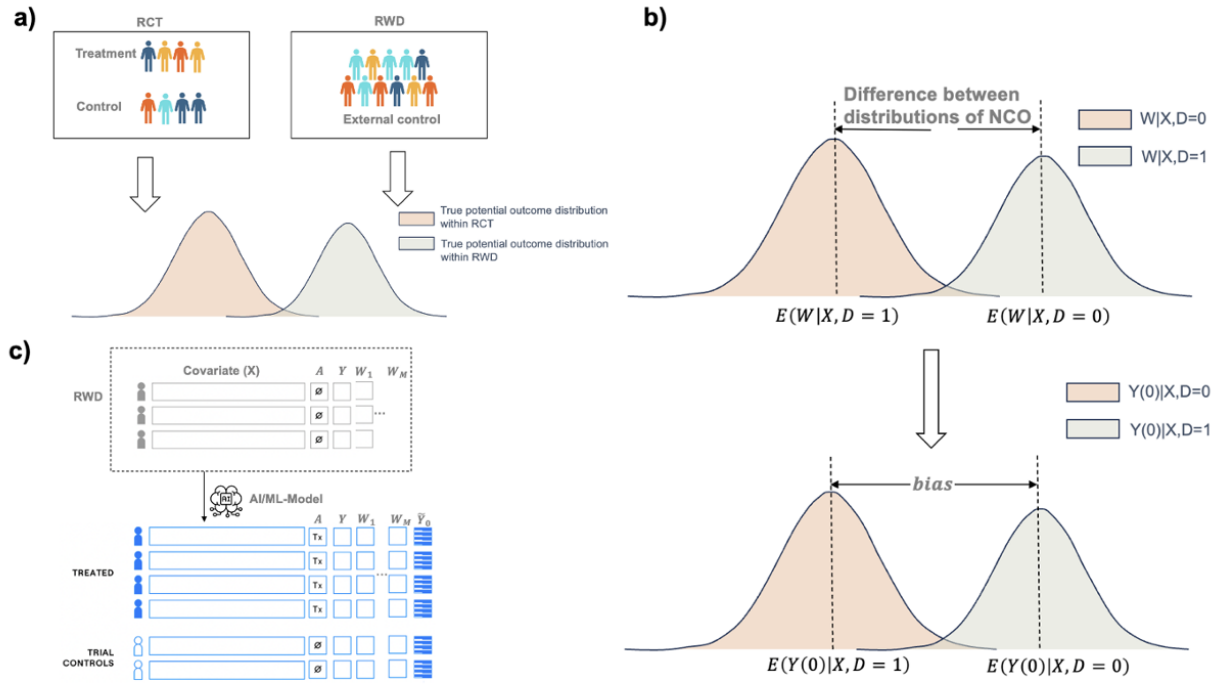

Figure 6.1: The working flow of the proposed method. a) the indication of the scenario, that is counterfactually, the outcome distribution given control for the RCT and RWD are different. The difference between the distribution of the outcome for the RCT and the RWD are known as the model-shift bias. b) negative control outcomes (NCOs) are those outcomes whose treatment response are hypothetically zero. The difference between the RCT and RWD for the NCOs helps to identify the model-shift for the outcomes of interest. c) The overview of the proposed method. The model used to predict the digital twin (shown in the last column for the RCT data) can be trained within the RWD data.

6.2. Results

Overview of Proposed Methods

Figure 6.1 (b) and (c) showed the overall working proposed method. Figure 6.1b introduces the concept of NCOs, outcomes that are theoretically unaffected by treatment assignment, allowing for an empirical estimate of the model shift bias. The observed discrepancy in NCO distributions between RCT and RWD serves as a diagnostic tool for identifying systematic shifts in outcome distributions. Finally, Figure 6.1c outlines our methodological framework: we utilize predictive modeling within the RWD to construct a digital twin—an individualized outcome model for RCT participants. More details about the proposed method can be found the Methods Section. By

systematically calibrating for model shift, our method facilitates more robust and generalizable treatment effect estimation in hybrid trial designs.

Empirical Evaluation of Proposed Method To evaluate the proposed method, we simulated 1,000 datasets combining RCT and RWD under two scenarios: (1) a linear relationship between features and outcomes and (2) a complex non-linear relationship, as described in Chen et al. (2015).

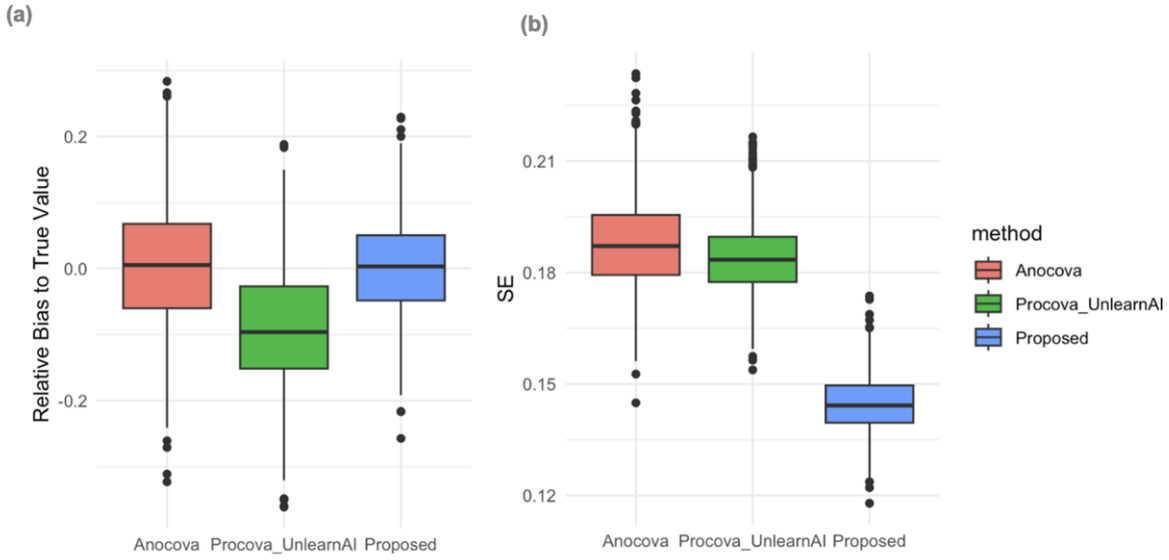

Figure 6.2: The results of the evaluation for synthetic data experiments, conducted with 1,000 iterations, are presented. The red boxplot represents the ANCOVA method, the green boxplot represents the PROCOVA method, and the blue boxplot represents the proposed method. (a) Relative bias compared to the true value across different methods. (b) Standard error across different methods.

Figure 6.2 illustrates the performance of the proposed method in comparison to baseline methods, including PROCOVA—an approach that does not account for model-shift bias—and ANCOVA. The ANCOVA method relies solely on RCT data, making its point estimates the gold standard due to the inherent randomization in RCTs. First, the proposed method produces estimates that are nearly identical to those from the ANCOVA method, with a relative bias of 10.2% compared to the true value, whereas the PROCOVA method has a relative bias of 0.2%. Second, the proposed method achieves a 39.9% relative efficiency gain compared to ANCOVA. In contrast, the PROCOVA

method, which is affected by model-shift bias, shows minimal efficiency improvement over ANCOVA. These results demonstrate that the proposed method effectively preserves the validity of ANCOVA while significantly improving efficiency. By addressing model-shift bias, it provides a more reliable approach than PROCOVA when integrating external real-world data with RCTs.

Combining SPRINT-MIND trial with iSTAGING Consortium

To evaluate the proposed method in real-world data, we applied it to a combined dataset from the SPRINT-MIND trial and the iSTAGING Consortium. The SPRINT-MIND trial investigated the effects of intensive blood pressure control on cerebral white matter lesions. Participants underwent magnetic resonance imaging (MRI) to assess changes in brain structure. The study found that intensive systolic blood pressure control was associated with a smaller increase in white matter lesion volume compared to standard treatment. Table 6.1 presents the baseline characteristics of participants in the SPRINT-MIND trial. The treatment group was defined as those receiving intensive antihypertensive therapy targeting a systolic blood pressure (SBP) of 120 mm Hg, while the control group received standard antihypertensive therapy targeting an SBP of 140 mm Hg.

The iSTAGING Consortium is an international collaboration that conducted studies across diverse geographic locations between 1995 and 2020, utilizing various scanners and acquisition parameters. For our analysis, we focused on cross-sectional scans from a subgroup of participants (n = 2,558) within the iSTAGING dataset who had complete MRI measures available. The iSTAGING Consortium integrates data from three sources: ADNI (n = 2,343), OASIS (n = 22), and Penn (n = 194). Baseline MRI data were missing for 633 participants (25.5%), leaving 1,905 participants for training the digital twin models. Key covariates for training the digital twins included age, sex, race, and education. The machine learning model used for digital twin construction was based on random forests. Table E.1 in the Supplement listed the NCO outcomes used for the analysis.

Efficiency gain for proposed method in case of SPRINT-MIND plus iSTAGING consortium

Figure 6.3 illustrates the estimated effect sizes across three brain regions—(a) deep cerebral white

Table 6.1: Summary-level characteristics for the SPRINT-MIND trial.

Characteristic	Standard Hypertension (N=200)	Intensive Hypertension (N=245)	Overall (N=445)
Sex			
Male	131 (65.5%)	148 (60.4%)	279 (62.7%)
Female	69 (34.5%)	97 (39.6%)	166 (37.3%)
Age			
Mean (SD)	66.4 (7.84)	67.8 (7.71)	67.2 (7.80)
Median [Min, Max]	65.0 [55.0, 89.0]	67.0 [55.0, 88.0]	66.0 [55.0, 89.0]
Cardiovascular history			
No	179 (89.5%)	216 (88.2%)	395 (88.8%)
Yes	21 (10.5%)	29 (11.8%)	50 (11.2%)
Race/ethnicity			
Non-Hispanic Black	65 (32.5%)	68 (27.8%)	133 (29.9%)
Hispanic	13 (6.5%)	9 (3.7%)	22 (4.9%)
Other	4 (2.0%)	2 (0.8%)	6 (1.3%)
Non-Hispanic White	118 (59.0%)	166 (67.8%)	284 (63.8%)
Myocardial infarction			
No	195 (97.5%)	236 (96.3%)	431 (96.9%)
Yes	5 (2.5%)	9 (3.7%)	14 (3.1%)
Stroke			
No	200 (100%)	244 (99.6%)	444 (99.8%)
Yes	0 (0%)	1 (0.4%)	1 (0.2%)
Acute coronary syndrome			
No	196 (98.0%)	238 (97.1%)	434 (97.5%)
Yes	4 (2.0%)	7 (2.9%)	11 (2.5%)
Now smoke			
No	79 (39.5%)	104 (42.4%)	183 (41.1%)
Yes	21 (10.5%)	37 (15.1%)	58 (13.0%)
Missing	100 (50.0%)	104 (42.4%)	204 (45.8%)
Education level			
Less than high school	16 (8.0%)	14 (5.7%)	30 (6.7%)
High school graduate	25 (12.5%)	40 (16.3%)	65 (14.6%)
Some college or associate degree	75 (37.5%)	81 (33.1%)	156 (35.1%)
College graduate or higher	80 (40.0%)	109 (44.5%)	189 (42.5%)
Missing	4 (2.0%)	1 (0.4%)	5 (1.1%)
Heart attack			
No	196 (98.0%)	235 (95.9%)	431 (96.9%)
Yes	4 (2.0%)	10 (4.1%)	14 (3.1%)
Depression			
No	172 (86.0%)	199 (81.2%)	371 (83.4%)
Yes	28 (14.0%)	46 (18.8%)	74 (16.6%)
BMI			
Mean (SD)	29.5 (5.05)	29.5 (5.18)	29.5 (5.12)
Median [Min, Max]	29.2 [18.2, 46.4]	28.8 [19.4, 45.3]	29.0 [18.2, 46.4]
Missing	1 (0.5%)	0 (0%)	1 (0.2%)

matter, (b) deep cerebral white matter left, and (c) deep cerebral white matter right—comparing the proposed method with two baseline approaches, PROCOVA and ANCOVA. As ANCOVA relies solely on RCT data, its estimates serve as the gold standard due to the benefits of randomization. The PROCOVA method incorporates external real-world data (RWD) but does not account for model-shift bias, leading to potential inaccuracies. The proposed method, by addressing model-shift bias, produces estimates that align closely with ANCOVA while achieving notable efficiency gains. Across all panels, the proposed method consistently yields more precise estimates, reflected in its narrower confidence intervals compared to ANCOVA, and avoids the potential distortions observed in PROCOVA.

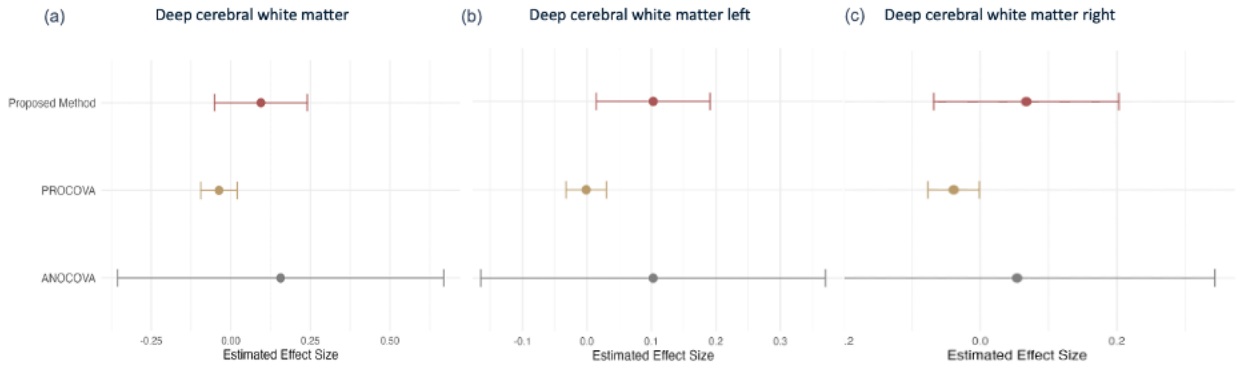

Figure 6.3: Estimated effect sizes and confidence intervals for three brain regions—(a) deep cerebral white matter, (b) deep cerebral white matter left, and (c) deep cerebral white matter right—using the proposed method (dark red), PROCOVA (gold), and ANCOVA (gray).

6.3. Discussion

The integration of RWD with RCTs has the potential to enhance trial efficiency, improve generalizability, and optimize clinical decision-making. However, systematic differences between RCT and RWD populations, known as model shift bias, pose significant challenges when leveraging external data sources. In this study, we propose a NCO-calibrated digital twin approach that systematically corrects for these discrepancies by quantifying and adjusting for systematic bias between trial and real-world settings. Our results demonstrate that this method preserves the internal validity of traditional ANCOVA while significantly enhancing statistical efficiency, thereby offering a robust

framework for integrating multi-source, multi-site RWD into clinical trials. Through extensive simulation studies and real-world applications, we show that the proposed method effectively reduces bias, improves treatment effect precision, and ensures robust causal inference in hybrid trial designs.

A key advantage of our approach lies in its ability to empirically adjust for bias using NCOs, rather than relying on untestable assumptions of data exchangeability between RCTs and RWD. By incorporating digital twin technology, our method constructs personalized counterfactual control outcomes that better account for heterogeneity in patient populations. These findings suggest that the digital twin framework, when properly calibrated, can serve as a powerful tool for leveraging real-world evidence in regulatory decision-making, comparative effectiveness research, and precision medicine applications. Future work will explore the application of this method across diverse therapeutic areas and broader patient populations, further refining its utility in real-world clinical research settings.

6.4. Methods

Training the Digital Twin Model Using External Real-World Data

To enhance the robustness and generalizability of treatment effect estimation in Randomized Controlled Trials (RCTs), we introduce a digital twin framework leveraging external Real-World Data (RWD). A digital twin is a model-derived representation of an individual's expected outcome under a counterfactual condition—in this case, the untreated outcome for trial participants.

Prognostic Modeling with External Real-World Data

We leverage external RWD to estimate individual-level prognostic outcomes under control conditions. Specifically, we train a predictive model using a machine learning algorithm (e.g., random forest) to estimate a prognostic function $\Psi(X \mid D = 0)$, where X represents observed covariates and $D = 0$ denotes untreated individuals from the external population. The prognostic function is defined as:

$$\Psi(X \mid D = 0) = E(Y(0) \mid X, D = 0),$$

where $Y(0)$ represents the outcome under control conditions.

This ensures that, conditional on $\Psi(X \mid D = 0)$, the outcome $Y(0)$ is independent of covariates X, reducing potential confounding. Random forests are particularly suitable due to their ability to capture complex, non-linear relationships in high-dimensional settings.

Negative-Control Outcome Calibration

Direct application of the digital twin model to the RCT population may introduce model-shift bias since distributions may differ between RWD and trial settings, expressed as:

$$P(Y_0 \mid X, D = 1) \neq P(Y_0 \mid X, D = 0).$$

To mitigate this, we introduce Negative-Control Outcome (NCO) calibration. NCOs are outcomes theoretically unaffected by treatment but share confounding structures with the primary outcome. Leveraging M NCOs, denoted W_m for $m = 1, \ldots, M$, we assume:

$$E(Y(0) \mid X, D = 1) - E(Y(0) \mid X, D = 0) = E(W \mid X, D = 1) - E(W \mid X, D = 0),$$

illustrated here with one NCO, W.

For each NCO W_m, we estimate two parallel functions:

$$\hat{\Psi}^{(0,m)}(X) = E(W_m \mid X, D = 0), \quad \hat{\Psi}^{(1,m)}(X) = E(W_m \mid X, D = 1).$$

The discrepancies between $\hat{\Psi}^{(0,m)}(X)$ and $\hat{\Psi}^{(1,m)}(X)$ calibrate the baseline digital twin function:

$$\tilde{\Psi}^0(X) = \hat{\Psi}^0(X) - \sum_{m=1}^{M} \left[\hat{\Psi}^{(0,m)}(X) - \hat{\Psi}^{(1,m)}(X) \right].$$

This calibrated function $\tilde{\Psi}^0(X)$ accounts for model-shift bias.

Generating Digital Twins within the RCT Population

After calibration, the refined digital twin generator provides a bias-adjusted control outcome for

each RCT participant i:

$$\tilde{Y}_{i,0} = \tilde{\Psi}^0(X_i).$$

This individual-level digital twin enhances reliability and reduces biases from model shift.

Treatment Effect Estimation using Digital Twin-Augmented ANCOVA

We estimate the causal treatment effect via ANCOVA incorporating the digital twin predictions:

$$Y_i = \beta_0 + \beta_1 A_i + \beta_\Psi \tilde{Y}_{i,0} + e_i,$$

where Y_i is the observed outcome, A_i the treatment indicator ($A_i = 1$ for treated, $A_i = 0$ for control), and e_i the residual error. The coefficient β_1 estimates the Average Treatment Effect (ATE).

Robust Uncertainty Estimation and Inference

We compute the standard error (SE) of β_1 using a sandwich estimator to provide heteroskedasticity-robust variance estimates, accounting for potential ANCOVA misspecification. This ensures robust inference for treatment effects despite model shifts and unmeasured confounding.

CHAPTER 7

CONCLUSION

This dissertation presents a cohesive and forward-looking framework to enhance the integration of RCTs with RWD. Through three major aims, I introduced methodological innovations that bridge the gap between the rigor of RCTs and the generalizability of RWD, ultimately empowering causal inference in diverse healthcare settings.

The first aim tackled the complexities of federated learning in decentralized data environments by developing a novel one-shot distributed algorithm, ODACoRH, for competing risks analysis. This approach not only preserved patient privacy but also improved efficiency and accuracy across heterogeneous clinical sites. The framework demonstrated superior performance to existing meta-analytic techniques, especially in settings with rare events, such as pediatric Crohn's disease and otitis media, thereby expanding the utility of distributed research networks.

The second aim introduced the Negative Control-Calibrated Difference-in-Differences (NC-DiD) framework within the RWD, addressing the problem of time-varying unmeasured confounding in observational studies. By leveraging NCOs, NC-DiD allowed for systematic detection and calibration of biases, improving the validity of causal inferences drawn from EHR. This methodology was effectively applied to study of pediatric long COVID, producing more robust and interpretable estimates compared to standard DiD models.

The third aim developed the Negative Control-Calibrated Digital Twin framework to integrate RCT and RWD populations, mitigating model shift bias. By calibrating individualized digital twins with NCOs, this approach enhanced statistical efficiency and preserved internal validity in clinical trials. The empirical validation using SPRINT-MIND and iSTAGING datasets demonstrated its promise for more inclusive and externally valid clinical trial designs.

Collectively, the contributions of this dissertation advance both theory and practice in biostatistics and digital health. Methodologically, the proposed frameworks extend existing causal inference tools

to accommodate the complexities of modern healthcare data, including site heterogeneity, unmeasured confounding, and population differences. Practically, the algorithms developed herein provide scalable and privacy-preserving solutions ready for implementation in clinical research networks and healthcare systems.

In conclusion, this work lays a foundation for a new era of methodological, data-driven, and inclusive clinical research. By embracing the complexity of multi-site, multi-source RWD and addressing key biases in observational studies, we move closer to realizing the full potential of learning health systems that are both evidence-based and patient-centered.

APPENDIX A

SUPPLEMENTARY APPENDIX FOR CHAPTER 2

A.1. Log-likelihood of Model for Competing Risk Data with Heterogeneity and Local-site Competing Risks Model

If data from different sites can be pooled together, the pooled likelihood for the k-th type ($k = 1, \ldots, K$) is

$$L_k(\beta_k) = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} 1(\delta_{ij}\pi_{ij} = k) \log \left(\frac{e^{\beta_k^T Z_{ij}}}{\sum_{s \in R_j(X_{ij})} \hat{w}_{ij,s} e^{\beta_k^T Z_s}} \right), \tag{A.1}$$

where $R_j(t)$ is the risk set defined as $R_j(t) = \{s : \{X_{sj} > t\} \cup \{X_{sj} \leq t \cap \Pi \neq k\}\}$ and $\hat{w}_{ij,s} = \frac{\hat{G}_j(X_{ij})}{\hat{G}_j(X_{sj} \wedge X_{ij})}$. $\hat{G}_j(\cdot)$ denotes the Kaplan-Meier estimator of the survival function of censoring from the j-th site ($j = 1, \ldots, J$).

The local-site data log-likelihood for the k-th type on the j-th site is

$$L_{jk}(\beta_k) = \frac{1}{n_j} \sum_{i=1}^{n_j} 1(\delta_{ij}\pi_{ij} = k) \log \left(\frac{e^{\beta_k^T Z_{ij}}}{\sum_{s \in R_j(X_{ij})} \hat{w}_{ij,s} e^{\beta_k^T Z_s}} \right), \tag{A.2}$$

where $R_j(t)$ is defined similarly, and the local estimate $\hat{\beta}_{jk}$ is obtained by

$$\hat{\beta}_{jk} = \underset{\beta_k}{\operatorname{argmax}}\, L_{jk}(\beta_k). \tag{A.3}$$

A.2. Formulas for $\nabla L_{1k}(\bar{\beta}_k)$, $\nabla^2 L_{1k}(\bar{\beta}_k)$, $\nabla L_k(\bar{\beta}_k)$, and $\nabla^2 L_k(\bar{\beta}_k)$

We need to calculate $\nabla L_{1k}(\bar{\beta}_k)$, $\nabla^2 L_{1k}(\bar{\beta}_k)$, $\nabla L_k(\bar{\beta}_k)$, and $\nabla^2 L_k(\bar{\beta}_k)$ without sharing patient-level data:

$$\nabla L_{1k}(\bar{\beta}_k) = \frac{1}{n_1} \sum_{i=1}^{n_1} I(\delta_{i1}\pi_{i1} = k) \left\{ Z_{i1} - \frac{\sum_{s \in R_1(x_{i1})} w_{i1,s} \exp(\bar{\beta}_k^T Z_s) Z_s}{\sum_{s \in R_1(x_{i1})} w_{i1,s} \exp(\bar{\beta}_k^T Z_s)} \right\}, \tag{A.4}$$

$$\nabla^2 L_{1k}(\bar{\beta}_k) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(\delta_{i1}\pi_{i1}=k)\left\{\frac{\sum_{s\in R_1(x_{i1})} w_{i1,s}\exp(\bar{\beta}_k^T Z_s) Z_s^{\otimes 2}}{\sum_{s\in R_1(x_{i1})} w_{i1,s}\exp(\bar{\beta}_k^T Z_s)} - \left[\frac{\sum_{s\in R_1(x_{i1})} w_{i1,s}\exp(\bar{\beta}_k^T Z_s) Z_s}{\sum_{s\in R_1(x_{i1})} w_{i1,s}\exp(\bar{\beta}_k^T Z_s)}\right]^{\otimes 2}\right\}. \quad \text{(A.5)}$$

$$\nabla L_k(\bar{\beta}_k) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} I(\delta_{ij}\pi_{ij}=k)\left\{Z_{ij} - \frac{\sum_{s\in R_j(x_{ij})} w_{ij,s}\exp(\bar{\beta}_k^T Z_s) Z_s}{\sum_{s\in R_j(x_{ij})} w_{ij,s}\exp(\bar{\beta}_k^T Z_s)}\right\}, \quad \text{(A.6)}$$

$$\nabla^2 L_k(\bar{\beta}_k) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} I(\delta_{ij}\pi_{ij}=k)\left\{\frac{\sum_{s\in R_j(x_{ij})} w_{ij,s}\exp(\bar{\beta}_k^T Z_s) Z_s^{\otimes 2}}{\sum_{s\in R_j(x_{ij})} w_{ij,s}\exp(\bar{\beta}_k^T Z_s)} - \left[\frac{\sum_{s\in R_j(x_{ij})} w_{ij,s}\exp(\bar{\beta}_k^T Z_s) Z_s}{\sum_{s\in R_j(x_{ij})} w_{ij,s}\exp(\bar{\beta}_k^T Z_s)}\right]^{\otimes 2}\right\}. \quad \text{(A.7)}$$

Equations (A.6) and (A.7) using summary-level information are:

$$\nabla L_k(\bar{\beta}_k) = \sum_{j=1}^{J} \nabla L_{jk}(\bar{\beta}_k), \quad \text{(A.8)}$$

$$\nabla^2 L_k(\bar{\beta}_k) = \sum_{j=1}^{J} \nabla^2 L_{jk}(\bar{\beta}_k). \quad \text{(A.9)}$$

A.3. Estimator of Variance for ODACoRH and Meta-analysis

The variance for ODACoRH is calculated as:

$$\hat{V}(\tilde{\beta}_k) = \frac{n_1}{n}\left[\nabla^2 L_{1k}(\tilde{\beta}_k)\right]^{-1}\left(\sum_{i=1}^{n_1}\nabla L_{1k}(\tilde{\beta}_k; x_{i1})\nabla L_{1k}(\tilde{\beta}_k; x_{i1})^T\right)\left[\nabla^2 L_{1k}(\tilde{\beta}_k)\right]^{-1}, \quad \text{(A.10)}$$

where

$$L_{1k}(\tilde{\beta}_k; x_{i1}) = 1(\delta_{i1}\pi_{i1} = k) \log \left(\frac{e^{\beta_k^T Z_{i1}}}{\sum_{s \in R_1(X_{i1})} \hat{w}_{i1,s} e^{\beta_k^T Z_s}} \right).$$

The variance for the meta-analysis method is calculated by:

$$\hat{V}(\bar{\beta}_k) = \left(\sum_{j=1}^{J} \hat{V}_{jk}^{-1} \right)^{-1}. \tag{A.11}$$

A.4. Simulation Studies and Figures

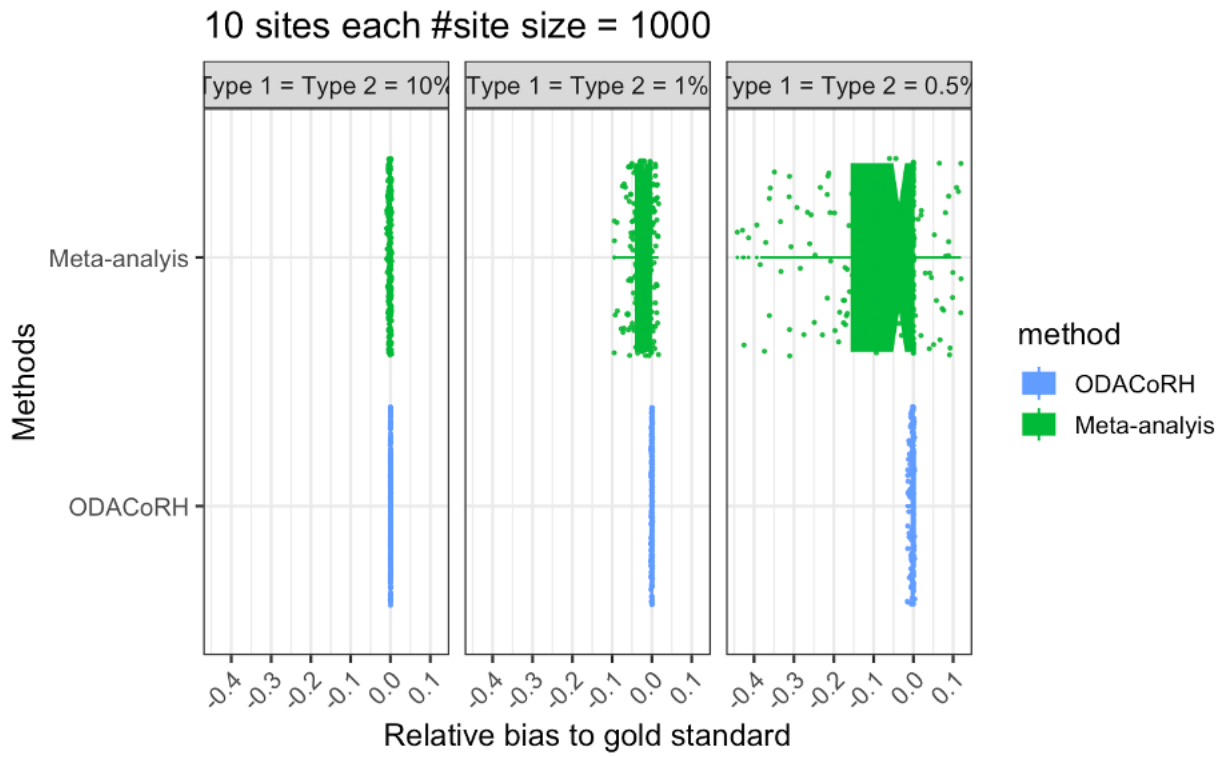

Figure A.1: Simulation results for the second component of β. Relative bias comparison between ODACoRH and meta-analysis under homogeneous conditions.

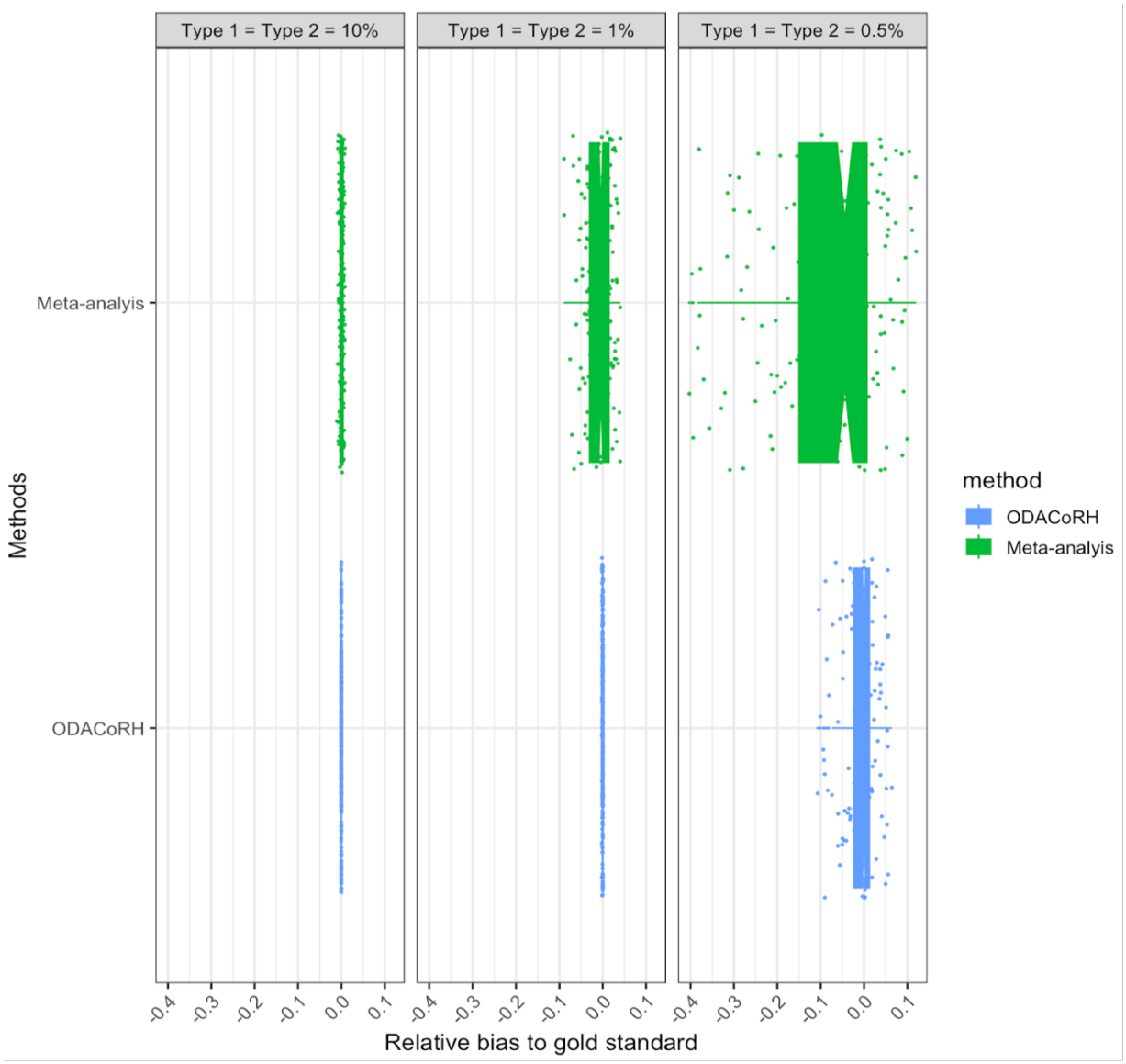

Figure A.2: Relative bias of ODACoRH and meta-analysis methods to the gold standard for the first component of the coefficient for type 1 event under homogeneous conditions.

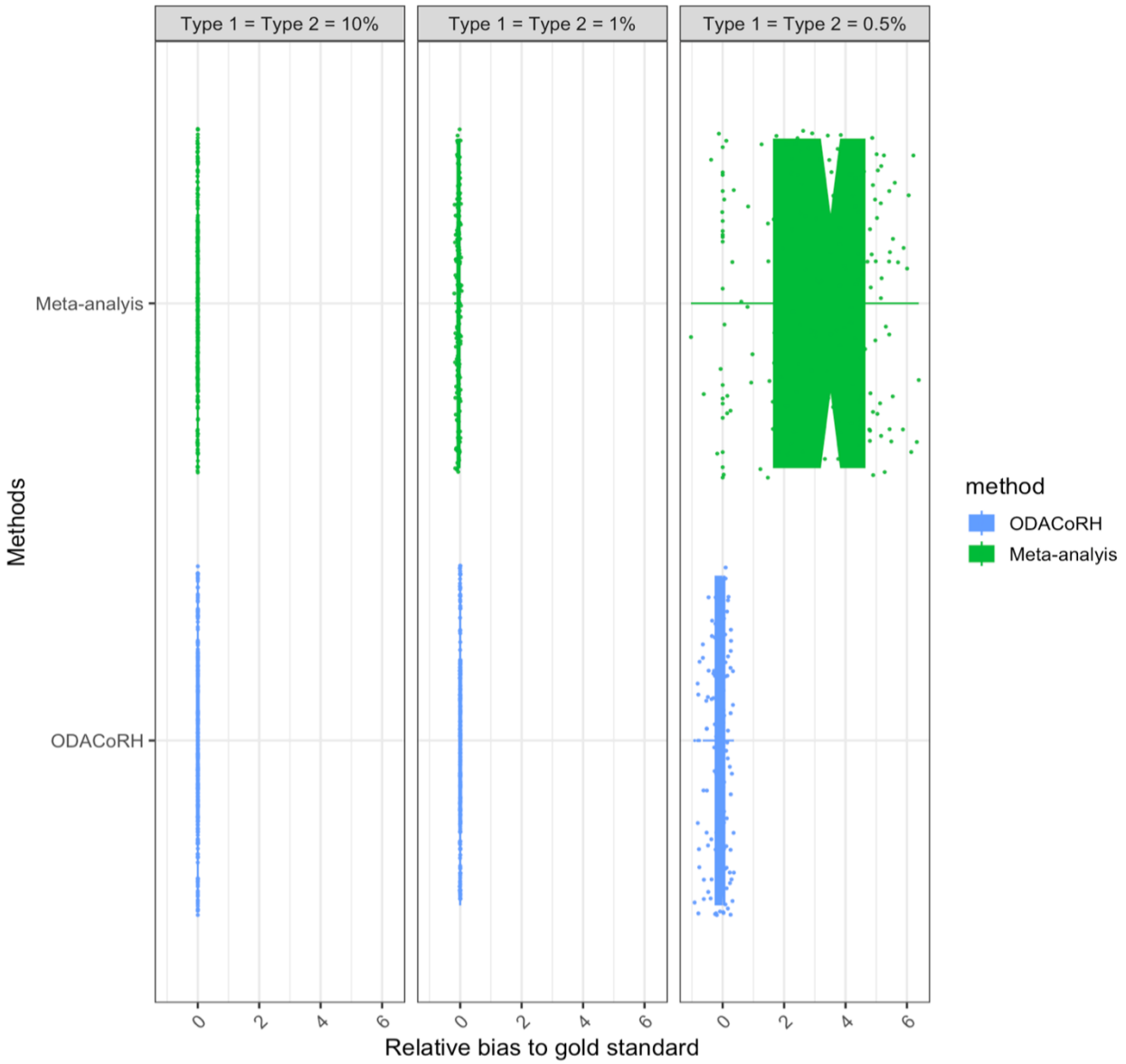

Figure A.3: Relative bias of ODACoRH and meta-analysis methods to the gold standard for the second component of the coefficient for type 1 event under homogeneous conditions.

A.5. Evaluation of Generalizability Using OM Study

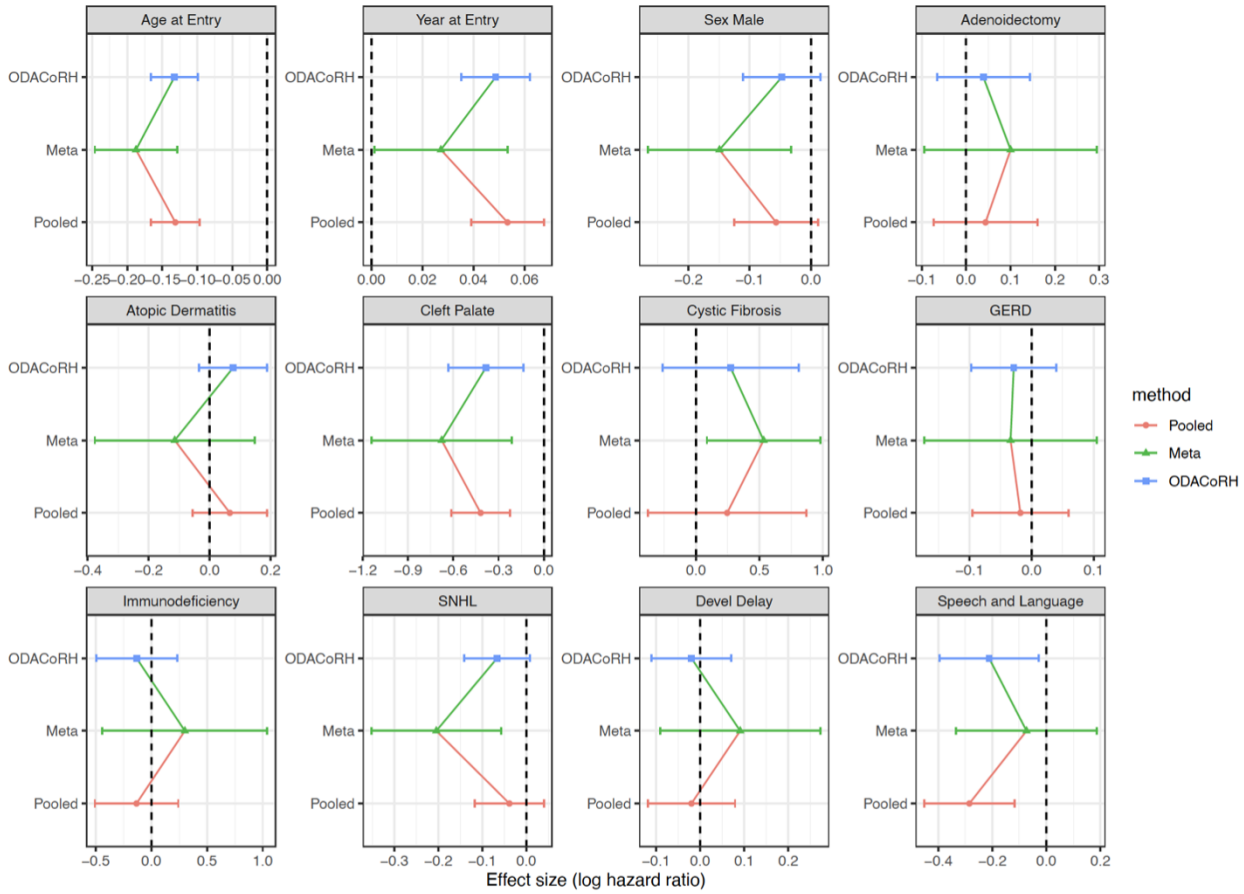

Figure A.4: Effect size of risk factors associated with recurrence of otitis media (OM). Effect sizes are in log-scale of subhazard ratios. Lines represent pooled (red), meta-analysis (green), and ODACoRH (blue) estimates. GERD indicates gastroesophageal reflux disease, and SNH indicates sensorineural hearing loss.

APPENDIX B

SUPPLEMENTARY APPENDIX FOR CHAPTER 3

B.1. Notation

We begin the proof by stating the technical lemmas, which has been stated in the federated learning literature. We first examined these lemmas under the rare disease setting and explicitly show the modifications of these lemmas under rare disease setting. To help the illustration, we define the following mathematical calculation form. We define the Orlicz norm of Z by $\|Z\|_{\psi_r} := \sup_{u \in S^{p-1}} \|u^{\mathrm{T}} Z\|_{\psi_r}$, where S^{p-1} denotes the unit sphere in $\mathbb{R}^p$. We define the following norm: $\|Z\|_q := (\sum_{i=1}^p |Z_i|^q)^{1/q}$ and $\|Z\|_\infty := \max_{i \in [p]} |Z_i|$. also define the operator norm of the kth order symmetric tensor $T \in \mathbb{R}^{p \times \ldots \times p}$ by $\|T\|_2 := \max_{\|u\|_2 = 1} \|T u^{\otimes(k-1)}\|_2$, where $u^{\otimes(k-1)} = u \otimes u \otimes \ldots \otimes u$. Also, we define the $tr(A)$ as trace of matrix A. Define A^2 as AA, the matrix product of the matrix A and itself. Define $a \vee b = \max(a, b)$. We also define the ith element of the tensor product for the tensor T and vector x as

$$(T x^{\otimes k-1})_i = \sum_{j_1, \ldots, j_{k-1}} T_{i, j_1, \ldots, j_{k-1}} \prod_{t=1}^{k-1} x_{j_t}.$$

We also use $\lesssim$ to denote the positive semi-definite ordering, for example, $A \lesssim B$ means that $A - B$ is a semi-definite positive matrix. We define the Frobenius norm of a $m \times n$ matrix A by

$$\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |A_{i,j}|^2 \right)^{1/2}.$$

B.2. summary statistics for $\nabla L_F(\tilde{\theta}^{(1)})$

From the main manuscript Equation (3), the formula for the surrogate Firth-corrected log-likelihood $\tilde{L}$ shows that $L_1(\tilde{\theta}^{(1)}$ can be found by the local-site data for the lead site. The problem is how to

construct $\nabla L(\tilde{\theta}^{(1)})$ with the summary-level statistic. We found

$$\nabla L_F(\tilde{\theta}^{(1)}) = \nabla L(\tilde{\theta}^{(1)}) + b(\tilde{\theta}^{(1)}),$$

where the rth element for the $b(\tilde{\theta}^{(1)})$ can be denoted as $b_r(\tilde{\theta}^{(1)}) = (nm)^{-1}\text{tr}\{I(\tilde{\theta}^{(1)})\partial I^{-1}(\tilde{\theta}^{(1)})/\partial\theta_r\}$.

The first part $\nabla L(\tilde{\theta}^{(1)})$ can be seen as the sum of the first-order derivative from each site, that is

$$\nabla L(\tilde{\theta}^{(1)}) = m^{-1}\sum_{j=1}^{m}\nabla L_j(\tilde{\theta}^{(1)}).$$

The second part $\nabla L(\tilde{\theta}^{(1)})$ can be decomposed as a function of the sum of the second or third-order derivative from each site,

$$b_r(\tilde{\theta}^{(1)}) = m^{-1}\text{tr}[\{\sum_{j=1}^{m}\nabla^2 L_j(\tilde{\theta}^{(1)})\}^{-1}\{\sum_{j=1}^{m}\nabla^3_{\theta_r} L_j(\tilde{\theta}^{(1)})\}].$$

B.3. technical lemmas

We state the following assumptions:

Assumption 1. We assume β_0 is fixed and α_0 goes to negative infinity. The vector Z, corresponding to the coefficient β_0, satisfies $E(Z) = 0$ and that $\|Z\|_{\psi_2} \leq K$ with $K \geq 1$. Besides, $\lambda_{\min}\left\{E(ZZ^T)\right\} > 0$.

Assumption 2. $\forall\eta \in \mathbb{R}$, $g(Z;\beta) \leq M$, $|g(Z;\beta)| \leq M$, and with $1 \leq M < \infty$.

Assumption 3. There exists positive number μ_- and μ_+ such that $\mu_- I_{p+1} \lesssim E\left\{\nabla^2 L(\theta;X)\right\} \lesssim \mu_+ I_{p+1}$.

From assumptions, we can define $\alpha_n = ng_1(\alpha_0)$ where $g_1(\alpha_0) = \exp(\alpha_n)$ and thus we have $\alpha_n = n_e$. Furthermore, we define $g_2(Z_i;\alpha_0) = \exp(Z_i^T\beta)$.

The assumption about parameters can be easy to find and is a standard regular condition. For

example,

$$E(\exp[t\left\{Y-h'(X^{\mathrm{T}}\theta)\right\}]|X)=\exp[\left\{h(X^{\mathrm{T}}\theta+\phi t)-h(X^{\mathrm{T}}\theta)-\phi t h'(X^{\mathrm{T}}\theta)\right\}/\phi]\leq\exp(\phi M t^2/2).$$

This implies that $\|Y-h'(X^{\mathrm{T}}\theta)\|_{\psi_2}\leq(M\phi)^{1/2}$. Moreover, the second order gradient has the property $\mu_- I_{p+1}\lesssim E\left\{\nabla^2 L(\theta;X)\right\}\lesssim\mu_+ I_{p+1}$.

Lemma B.3.1. *Under Assumptions 1 and 3, for jth site $(j=1,...,m)$ and for any $\xi>0$ and any $\theta\in\mathbf{R}^{p+1}$, when $n\geq p^2\vee(\log n)^2$, we have with probability at least $1-2e^{\xi}$ that*

$$\|\nabla L_j(\theta)-E\left\{\nabla L_j(\theta)\right\}\|_2\leq(\phi M)^{1/2}K[\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2}+\frac{\{(p+1)\vee\xi\}}{\alpha_n}].$$

Proof. Without loss of generality, we let $j=1$ to track the index notation. We start the proof by finding the limit of sequence $\nabla L_1(\theta)$ as $n\to\infty$ and $g_1(\alpha)\to 0$ such that $ng_1(\alpha)\to\infty$. Notice that

$$\begin{aligned}
E\left\{\nabla L_1(\theta)\right\} &= E\left\{\alpha_n^{-1}\sum_{i=1}^{n}\nabla L_1(\theta;X_{i1})\right\}\\
&= \alpha_n^{-1}\sum_{i=1}^{n}E(X_{i1}[Y_{i1}-\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\left\{1+g_1(\alpha)g_2(Z_{i1};\beta)\right\}^2}])\\
&= \alpha_n^{-1}\sum_{i=1}^{n}E(X_{i1}[E(Y_{i1}|X_{i1})-\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\left\{1+g_1(\alpha)g_2(Z_{i1};\beta)\right\}^2}])\\
&= \alpha_n^{-1}\sum_{i=1}^{n}E(X_{i1}[\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\left\{1+g_1(\alpha)g_2(Z_{i1};\beta)\right\}^2}-\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\left\{1+g_1(\alpha)g_2(Z_{i1};\beta)\right\}^2}])\\
&= 0.
\end{aligned}$$

We show that the sequence of $\nabla L_1(\theta)$ is related to α_n and approach its expectation, which is 0. The rest of proof can give the details about the error bound of this sequence.

$$\begin{aligned}
\|\nabla L_1(\theta)\| &\leq 2\sup_{v\in N(1/2)}|<\nabla L_1(\theta),v>|\\
&\leq 2\sup_{v\in N(1/2)}|<\alpha_n^{-1}\sum_{i=1}^{n}X_{i1}[Y_{i1}-\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\left\{1+g_1(\alpha)g_2(Z_{i1};\beta)\right\}^2}],v>|
\end{aligned}$$

By v's sub-Gaussian property, we have

$$\|[Y_{i1}-\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}](X_{i1}^T v)\|_{\psi_1} \leq \|[Y_{i1}-\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}]\|_{\psi_2}\|X_{i1}^{\mathrm{T}}v\|_{\psi_2} \lesssim (\phi M)^{1/2}K.$$

By Bernstein's inequality, we have for any $t > 0$,

$$\begin{aligned}\Pr\left(\left|\alpha_n^{-1}\sum_{i=1}^n X_{i1}\left[Y_{i1}-\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}\right]\right| \right. &\geq \left. 2(\phi M)^{1/2}K\left\{\left(\frac{t}{\alpha_n}\right)^{1/2}+\frac{t}{\alpha_n}\right\}\right) \\ &\leq 2\exp\left\{-(t-p\log 6)\right\}.\end{aligned}$$

Let $\xi = t - p\log 6$ into the bound, we have with the probability at least $1-2\exp(-\xi)$ such that

$$\|\nabla L_1(\theta)\|_2 \leq (\phi M)^{1/2}K[\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2}+\frac{\{(p+1)\vee\xi\}}{\alpha_n}]$$

□

Lemma B.3.2. *Under Assumptions 1 and 3, for jth site $(j = 1, ..., m)$ and for any $\xi > 0$ and any $\theta \in \mathbf{R}^{p+1}$, we have with probability at least $1 - 2e^{\xi}$ that*

$$\|\nabla^2 L_j(\theta) - E\left\{\nabla^2 L_j(\theta)\right\}\|_2 \leq MK^2[\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2}+\frac{(p+1)\vee\xi}{\alpha_n}]$$

Proof. We begin the proof and let $j = 1$. Notice that

$$\begin{aligned}\nabla^2 L_1(\theta) &= \alpha_n^{-1}\sum_{i=1}^n \nabla^2 L_1(\theta; X_{i1}) \\ &= \alpha_n^{-1}g_1(\alpha)\sum_{i=1}^n \frac{g_2(Z;\beta)}{\{1+g_1(\alpha)g_2(Z;\beta)\}^2}X_{i1}X_{i1}^T \\ &= n^{-1}\sum_{i=1}^n \frac{g_2(Z;\beta)}{\{1+g_1(\alpha)g_2(Z;\beta)\}^2}X_{i1}X_{i1}^T \\ &= E\left\{g_2(Z;\beta)X_{i1}X_{i1}^T\right\}+o_p(1).\end{aligned}$$

The last equality is found by Lemma 28 in Wang (2019). We show that the sequence of $\nabla^2 L_1(\theta)$

is related to α_n and approaches to $E[g_2(Z;\beta)X_{i1}X_{i1}^T]$. The rest of proof can give the details about the error bound of this sequence, which can be found by using the similar arguments from Lemma B2 in Wang and Zhu (2022).

Then, we have

$$\begin{aligned}
\|\nabla^2 L_1(\theta) - E\left\{\nabla^2 L_1(\theta)\right\}\|_2 &= \max_{\|u\|_2=1} < [\nabla^2 L_1(\theta) - E\left\{\nabla^2 L_1(\theta)\right\}]u, u > \\
&\leq 2 \max_{u\in N(1/4)} < [\nabla^2 L_1(\theta) - E\left\{\nabla^2 L_1(\theta)\right\}]u, u > \\
&\leq 2 \max_{u\in N(1/4)} |\frac{1}{\alpha_n}\sum_{i=1}^{n} \frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}(X_{i1}^{\mathrm{T}}u)^2 \\
&\quad -E\left\{g_2(Z;\beta)(X_{i1}^{\mathrm{T}}u)^2\right\}|
\end{aligned}$$

From Lemma 2.7.6 in Vershynin (2010), we have

$$\begin{aligned}
&\|\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}(X_{i1}^{\mathrm{T}}u)^2 - E\left\{g_2(Z;\beta)(X_{i1}^{\mathrm{T}}u)^2\right\}\|_{\psi_1} \\
\leq\ & \|g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}u)^2 - E\left\{g_2(Z;\beta)(X_{i1}^{\mathrm{T}}u)^2\right\}\|_{\psi_1} \\
\leq\ & C_1 M\|(X_{i1}^{\mathrm{T}}u)^2\|_{\psi_1} \\
\leq\ & C_1 M\|X_{i1}^{\mathrm{T}}u\|_{\psi_2}^2 \\
\leq\ & C_1 M K^2,
\end{aligned}$$

where C_1 is constant.

By Bernstein's inequality, we have

$$
\begin{aligned}
& pr[\max_{u\in N(1/4)} |\alpha_n^{-1}\sum_{i=1}^{n}\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}(X_{i1}^{\mathrm{T}}u)^2 - E\left\{g_2(Z;\beta)(X_{i1}^{\mathrm{T}}u)^2\right\}| \\
& \geq C_2MK^2\max\left\{(p+1)^{1/2}/\alpha_n, t/\alpha_n\right\}] \\
\leq & \sum_{u\in N(1/4)} pr[\max_{u\in N(1/4)} |\alpha_n^{-1}\sum_{i=1}^{n}\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}(X_{i1}^{\mathrm{T}}u)^2 - E\left\{g_2(Z;\beta)(X_{i1}^{\mathrm{T}}u)^2\right\}| \\
& \geq C_2MK^2\max\left\{(p+1)^{1/2}/\alpha_n, t/\alpha_n\right\}] \\
\leq & 2\exp[-\{t-9\log(p+1)\}].
\end{aligned}
$$

With $\xi = t - 9\log(p+1)$, we can find a universal constant C_3 with probability at least $1-2\exp(\xi)$ and we have

$$
\|\nabla^2 L_1(\theta) - E\left\{\nabla^2 L_1(\theta)\right\}\|_2 \leq MK^2[\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2} + \frac{(p+1)\vee\xi}{\alpha_n}].
$$

□

Lemma B.3.3. *Under Assumptions 1 and 3, for any $\xi > 0$ and any $\theta \in \mathbf{R}^{p+1}$, for jth site $(j = 1, ..., m)$, we have with probability at least $1 - 2e^{\xi}$ that*

$$
\|\nabla^3 L_j(\theta) - E\left\{\nabla^3 L_j(\theta)\right\}\|_2 \leq MK^3[\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2} + \frac{\{(p+1)\vee\xi\}^{3/2}}{\alpha_n}]
$$

Proof. Let $j = 1$ without loss of generality and we start the proof by finding $E\left\{\nabla^3 L_1(\theta)\right\}$. Notice

that

$$\begin{aligned}
\nabla^3 L_1(\theta) &= \alpha_n^{-1}\sum_{i=1}^{n}\nabla^3 L_1(\theta; X_{i1}) \\
&= -\alpha_n^{-1}\sum_{i=1}^{n}[\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2} - \frac{2g_1^2(\alpha)g_2^2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^3}]X_{i1}\otimes X_{i1}\otimes X_{i1} \\
&= -n^{-1}\sum_{i=1}^{n}[\frac{g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2} - \frac{2g_1(\alpha)g_2^2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^3}]X_{i1}\otimes X_{i1}\otimes X_{i1} \\
&= -n^{-1}\sum_{i=1}^{n}\frac{g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}X_{i1}\otimes X_{i1}\otimes X_{i1} + o_p(1) \\
&= -E[g_2(Z;\beta)X_{i1}\otimes X_{i1}\otimes X_{i1}] + o_p(1).
\end{aligned}$$

The second last equality is by assumption, $g_1(\alpha)\to 0$ as $n\to\infty$. The last equality is found by Lemma 28 Wang (2019). We show that the sequence of $\nabla^3 L_1(\theta)$ is related to α_n and approaches to $E[g_2(Z;\beta)X_{i1}\otimes X_{i1}\otimes X_{i1}]$. The rest of proof can give the details about the error bound of this sequence, which can be found by using the similar arguments from Lemma B3 in Wang and Zhu (2022).

$$\begin{aligned}
\|\nabla^3 L_1(\theta) - E\left\{\nabla^3 L_1(\theta)\right\}\| &\le 2\sup_{v\in N(1/8)}|[\nabla^3 L_1(\theta) - E\left\{\nabla^3 L_1(\theta)\right\}]v\otimes v\otimes v| \\
&\le 2\sup_{v\in N(1/8)}|\alpha_n^{-1}\sum_{i=1}^{n}[\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2} - \frac{2g_1(\alpha)^2g_2^2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^3}] \\
&\quad (X_{i1}^{\mathrm{T}}v)^3 - E\left\{g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}v)^3\right\}|
\end{aligned}$$

From the nature that v follows sub-Gaussian distribution, we have

$$E(\exp[\left\{\frac{|g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}v)^3|}{MK^3}\right\}^{2/3}]) \le 2$$

This gives us $\|g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}v)^3\|_{\psi_{2/3}} \le MK^3$. By using the bound from Equation (3.6) in Adamczak et al.

(2011), we have for any $t > 0$.

$$pr(|\alpha_n^{-1}\sum_{i=1}^{n}[\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2} - \frac{2g_1(\alpha)^2 g_2^2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^3}](X_{i1}^{\mathrm{T}}v)^3 - E\left\{g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}v)^3\right\}| \\ \geq C_1 M K^3 \left\{(t/\alpha_n)^{1/2} + t^{3/2}/\alpha_n\right\}) \leq 2\exp\left\{-(t-3)\right\}.$$

where C_1 is a constant.

With the help of bound implied by $v \in N(1/8)$, we have

$$\begin{aligned} & pr[\|\nabla^3 L_1(\theta) - E\left\{\nabla^3 L_1(\theta)\right\}\|_2 \geq C_2 M K^3 \left\{(t/\alpha_n)^{1/2} + t^{3/2}/\alpha_n\right\}] \\ \leq\ & pr[\sup_{\|v\|_2=1} |\alpha_n^{-1}\sum_{i=1}^{n}[\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2} - \frac{2g_1(\alpha)^2 g_2^2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^3}] \\ & (X_{i1}^{\mathrm{T}}v)^3 - E\left\{g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}v)^3\right\}| \geq C_2 M K^3 \left\{(t/\alpha_n)^{1/2} + t^{3/2}/\alpha_n\right\}] \\ \leq\ & 2\exp\left\{-(t-3-p\log 17)\right\} \end{aligned}$$

Let $\xi = t - 3 - p\log 17$ into the bound, we have with the probability at least $1 - 2\exp(-\xi)$ such that

$$\|\nabla^3 L_1(\theta) - E\left\{\nabla^3 L_1(\theta)\right\}\|_2 \leq M K^3 [\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2} + \frac{\{(p+1)\vee\xi\}^{3/2}}{\alpha_n}]$$

□

Lemma B.3.4. *Under Assumptions 1 and 3, for jth site* $(j = 1, ..., m)$ *and for any* $\xi > 0$ *and any* $\theta \in \mathbf{R}^{p+1}$*, we have with probability at least* $1 - 2e^{\xi}$ *that*

$$\|\nabla^4 L_j(\theta) - E\left\{\nabla^4 L_j(\theta)\right\}\|_2 \leq M K^4 [\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2} + \frac{\{(p+1)\vee\xi\}^2}{\alpha_n}]$$

Proof. We let $j = 1$ without loss of generality and start the proof by finding the limit of sequence

$\nabla^4 L_1(\theta)$ as $n \to \infty$ and $g_1(\alpha) \to 0$ such that $ng_1(\alpha) \to \infty$. Notice that

$$\begin{aligned}
\nabla^4 L_1(\theta) &= \alpha_n^{-1} \sum_{i=1}^{n} \nabla^4 L_1(\theta; X_{i1}) \\
&= -\alpha_n^{-1} \sum_{i=1}^{n} [\frac{g_1(\alpha) g_2(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^2} - \frac{6 g_1^2(\alpha) g_2^2(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^3} + \frac{6 g_1^3(\alpha) g_2^3(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^4}] \\
&\quad X_{i1} \otimes X_{i1} \otimes X_{i1} \otimes X_{i1} \\
&= -n^{-1} \sum_{i=1}^{n} [\frac{g_2(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^2} - \frac{6 g_1(\alpha) g_2^2(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^3} + \frac{6 g_1^2(\alpha) g_2^3(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^4}] \\
&\quad X_{i1} \otimes X_{i1} \otimes X_{i1} \otimes X_{i1} \\
&= -n^{-1} \sum_{i=1}^{n} \frac{g_2(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^2} X_{i1} \otimes X_{i1} \otimes X_{i1} \otimes X_{i1} + o_p(1) \\
&= -E[g_2(Z;\beta) X_{i1} \otimes X_{i1} \otimes X_{i1} \otimes X_{i1}] + o_p(1).
\end{aligned}$$

We show that the sequence of $\nabla^4 L_1(\theta)$ is related to α_n and approaches to $E[g_2(Z;\beta) X_{i1} \otimes X_{i1} \otimes X_{i1} \otimes X_{i1}]$. The rest of proof can give the details about the error bound of this sequence, which can be found by using the similar arguments from Lemma B4 in Wang and Zhu (2022).

$$\begin{aligned}
\| \nabla^4 L_1(\theta) - E\left\{\nabla^4 L_1(\theta)\right\} \| &\leq 2 \sup_{v \in N(1/16)} |[\nabla^4 L_1(\theta) - E\left\{\nabla^4 L_1(\theta)\right\}] v \otimes v \otimes v \otimes v| \\
&\leq 2 \sup_{v \in N(1/16)} |\alpha_n^{-1} \sum_{i=1}^{n} [\frac{g_1(\alpha) g_2(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^2} - \frac{6 g_1^2(\alpha) g_2^2(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^3} + \\
&\quad \frac{6 g_1^3(\alpha) g_2^3(Z_{i1};\beta)}{\{1 + g_1(\alpha) g_2(Z_{i1};\beta)\}^4}] (X_{i1}^{\mathrm{T}} v)^4 - E\left\{g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}} v)^4\right\}|
\end{aligned}$$

By v's sub-Gaussian property, we have

$$E(\exp[\left\{\frac{|g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}} v)^4|}{MK^4}\right\}^{1/2}]) \leq 2$$

This gives us $\|g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}} v)^4\|_{\psi_{1/2}} \leq MK^4$. By using the bound from Equation (3.6) in Adamczak et al.

(2011), we have for any $t > 0$,

$$pr(|\alpha_n^{-1}\sum_{i=1}^{n}[\frac{g_1(\alpha)g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2} - \frac{6g_1^2(\alpha)g_2^2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^3} + \frac{6g_1^3(\alpha)g_2^3(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^4}](X_{i1}^{\mathrm{T}}v)^4$$
$$-E\left\{g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}v)^4\right\}|) \geq C_1MK^4\left\{(t/\alpha_n)^{1/2} + t^2/\alpha_n\right\}) \leq 2\exp\left\{-(t-3)\right\}.$$

where C_1 is a constant.

With the help of bound implied by $v \in N(1/16)$, we have

$$\begin{aligned} & pr[\|\nabla^4 L_1(\theta) - E\left\{\nabla^4 L_1(\theta)\right\}\|_2 \geq C_2MK^4\left\{(t/\alpha_n)^{1/2} + t^2/\alpha_n\right\}] \\ \leq \quad & 2\exp\left\{-(t-3-p\log 33)\right\} \end{aligned}$$

Let $\xi = t - 3 - p\log 33$ into the bound, we have with the probability at least $1 - 2\exp(-\xi)$ such that

$$\|\nabla^4 L_1(\theta) - E\left\{\nabla^4 L_1(\theta)\right\}\|_2 \leq MK^4[\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2} + \frac{\{(p+1)\vee\xi\}^2}{\alpha_n}]$$

□

Lemma B.3.5. *Under Assumptions 1 and 2, for any $\xi > 0$ and any $\theta \in \mathbf{R}^{p+1}$, we have with probability at least $1 - 2e^{\xi}$ that*

$$\|\nabla^4 L_j(\theta)\|_2 \leq MK^4[1 + \left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2} + \frac{\{(p+1)\vee\xi\}^2}{\alpha_n}]$$

Proof. From the Lemma B.3.4, we have with the probability at least $1 - 2\exp(-\xi)$ such that

$$\|\nabla^4 L_1(\theta) - E\left\{\nabla^4 L_1(\theta)\right\}\|_2 \leq MK^4[\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2} + \frac{\{(p+1)\vee\xi\}^2}{\alpha_n}].$$

Notice that $E\left\{g_2(Z_{i1};\beta)(X_{i1}^{\mathrm{T}}v)^4\right\} \leq MK^4\|\beta\|_2$, and

$$\|\nabla^4 L_1(\beta)\|_2 \leq 2\sup_{v\in N(1/16)}\|\alpha_n^{-1}\sum_{i=1}^{n}\frac{g_2(Z_{i1};\beta)}{\{1+g_1(\alpha)g_2(Z_{i1};\beta)\}^2}(X_{i1}^{\mathrm{T}}v)^4\|.$$

Then we have the results. □

Lemma B.3.6. *Under Assumption 1 to 3, for jth site, $n \geq C(p+1)^2 \vee \log^2 n$, with the following event holding*

$$\varepsilon = \left\{ \max_{\|u\|_2=1} \frac{1}{n} \sum_{i=1}^{n} (X_{ij}^{\mathrm{T}} u)^4 \leq CK^4 \right\},$$

for a constant C, we have

$$\begin{aligned}
\left\|\nabla^2 L_j(\theta_1) - \nabla^2 L_j(\theta_2)\right\|_2 &\lesssim MK^3 \|\theta_1 - \theta_2\|_2, \\
\left\|\nabla^3 L_j(\theta_1) - \nabla^3 L_j(\theta_2)\right\|_2 &\lesssim MK^{7/2} \|\theta_1 - \theta_2\|_2, \\
\left\|\nabla^4 L_j(\theta_1) - \nabla^4 L_j(\theta_2)\right\|_2 &\lesssim MK^4 \|\theta_1 - \theta_2\|_2.
\end{aligned}$$

for any θ_1 and θ_2.

Proof. We let $j = 1$ without loss of generality, and begin the proof for ∇L^2

$$\begin{aligned}
\|\nabla^2 L_1(\theta_1) - \nabla^2 L_1(\theta_2)\|_2 &= \| \frac{1}{\alpha_n} g_1(\alpha_1) \sum_{i=1}^{n} \frac{g_2(Z_{i1};\beta_1)}{\{1 + g_1(\alpha_1) g_2(Z_{i1};\beta_2)\}^2} X_{i1} X_{i1}^{\mathrm{T}} \\
&\quad - \frac{1}{\alpha_n} g_1(\alpha_2) \sum_{i=1}^{n} \frac{g_2(Z_{i1};\beta_2)}{\{1 + g_1(\alpha_2) g_2(Z_{i1};\beta_2)\}^2} X_{i1} X_{i1}^{\mathrm{T}} \|_2 \\
&\leq \| \frac{1}{n} \sum_{i=1}^{n} g_2(Z_{i1};\beta_1) X_{i1} X_{i1}^{\mathrm{T}} - \frac{1}{n} \sum_{i=1}^{n} g_2(Z_{i1};\beta_2) X_{i1} X_{i1}^{\mathrm{T}} \|_2 \\
&= \| \int_0^1 [\frac{1}{n} \sum_{i=1}^{n} g_2 \left\{ Z_{i1}; \beta_1 + v(\beta_2 - \beta_1) \right\} Z_{i1}^T (\beta_1 - \beta_2) X_{i1} X_{i1}^T] dv \|_2 \\
&\leq \| \int_0^1 [\frac{1}{n} \sum_{i=1}^{n} g_2 \left\{ Z_{i1}; \beta_1 + v(\beta_2 - \beta_1) \right\} X_{i1} \otimes X_{i1} \otimes X_{i1}] dv \|_2 \|\theta_1 - \theta_2\|_2,
\end{aligned}$$

where the last equality is from the Taylor's Expansion. We focus on the first part, which has

$$
\begin{aligned}
& \| \int_0^1 \frac{1}{n} \sum_{i=1}^{n} g_2 \left\{ Z_{i1}; \beta_1 + v(\beta_2 - \beta_1) \right\} X_{i1} \otimes X_{i1} \otimes X_{i1} dv \|_2 \\
\leq & \max_{\|u\|_2=1} | \int_0^1 \frac{1}{n} \sum_{i=1}^{n} g_2 \left\{ Z_{i1}; \beta_1 + v(\beta_2 - \beta_1) \right\} (X_{i1}^{\mathrm{T}} u)^3 dv | \\
\leq & \max_{\|u\|_2=1} \int_0^1 | \frac{1}{n} \sum_{i=1}^{n} g_2 \left\{ Z_{i1}; \beta_1 + v(\beta_2 - \beta_1) \right\} (X_{i1}^{\mathrm{T}} u)^3 | dv \\
\leq & \max_{\|u\|_2=1} \int_0^1 [\frac{1}{n} \sum_{i=1}^{n} g_2 \left\{ Z_{i1}; \beta_1 + v(\beta_2 - \beta_1) \right\} (X_{i1}^{\mathrm{T}} u)^2]^{1/2} [\frac{1}{n} \sum_{i=1}^{n} (X_{i1}^{\mathrm{T}} u)^4]^{1/2} dv \\
\leq & M \max_{\|u\|_2=1} \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{i1}^{\mathrm{T}} u)^2 \right\}^{1/2} \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{i1}^{\mathrm{T}} u)^4 \right\}^{1/2} \\
\leq & M \max_{\|u\|_2=1} \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{i1}^{\mathrm{T}} u)^4 \right\}^{3/4} \\
\leq & CMK^3
\end{aligned}
$$

With the similar arguments, we can prove that $\|\nabla^3 L_j(\theta_1) - \nabla^3 L_j(\theta_2)\|_2 \lesssim MK^{7/2} \|\theta_1 - \theta_2\|_2$ and $\|\nabla^4 L_j(\theta_1) - \nabla^4 L_j(\theta_2)\|_2 \lesssim MK^4 \|\theta_1 - \theta_2\|_2$. □

Lemma B.3.7. *Under Assumptions 1 to 3, for the jth site ($j = 1, .., m$), the local estimator $\hat{\theta}_j$ has the probability $1 - 4\exp(\xi) - 2n^{-8}$ that*

$$
\|\hat{\theta}_j - \theta_0\|_2 \leq C_1 \mu_-^{-1} (\phi M)^{1/2} K \left\{ \frac{(p+1) \vee \xi}{\alpha_n} \right\}^{1/2}.
$$

Proof. We begin to define the following events,

$$
\begin{aligned}
\varepsilon_{F,0} &= \left\{ \max_{\|u\|_2=1} \frac{1}{n} \sum_{i=1}^{n} (X_{i1}^{\mathrm{T}} u)^4 \leq C_1 K^4 \right\}, \\
\varepsilon_{F,1} &= \left\{ \|\nabla^2 L_{F,j}(\theta_0) - E\left\{ \nabla^2 L_{F,j}(\theta_0) \right\}\|_2 \leq \frac{\mu_-}{4} \right\}, \\
\varepsilon_{F,2} &= \left\{ \|\nabla L_{F,j}(\theta_0) - E\left\{ \nabla L_{F,j}(\theta_0) \right\}\|_2 \leq \frac{\mu_-^2 K}{32MK^3} \right\}.
\end{aligned}
$$

By applying the Lemma 6 from Zhang et al. (2012), we have

$$\|\hat{\theta}_j - \theta_0\|_2 \leq 2\mu_-^{-1}\|\nabla L_1(\theta_0) - E\left\{\nabla L_1(\theta_0)\right\}\|_2.$$

By applying the Lemma B.3.1, we have

$$\|\hat{\theta}_j - \theta_0\|_2 \leq C_1\mu_-^{-1}(\phi M)^{1/2}K\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2}.$$

□

Lemma B.3.8. *Under Assumptions 1 to 3, the local estimator* $\hat{\theta}_{F,j}$ *has the probability* $1-10\exp(-\xi)-2n^{-8}$ *that*

$$\|\hat{\theta}_{F,j} - \theta_0\|_2 \leq C_1\mu_-^{-1}(\phi M)^{1/2}K\left\{\frac{(p+1)\vee\xi}{\alpha_n}\right\}^{1/2}.$$

Proof. Without loss of generality, we let $j = 1$ and follow the proof of Lemma 6 from Zhang et al. (2012). We first define the following events,

$$\begin{aligned}
\varepsilon_{F,0} &= \left\{\max_{\|u\|_2=1}\frac{1}{n}\sum_{i=1}^{n}(X_{i1}^{\mathrm{T}}u)^4 \leq C_1K^4\right\},\\
\varepsilon_{F,1} &= \left\{\|\nabla^2 L_{F,j}(\theta_0) - E\left\{\nabla^2 L_{F,j}(\theta_0)\right\}\|_2 \leq \frac{\mu_-}{4}\right\},\\
\varepsilon_{F,2} &= \left\{\|\nabla L_{F,j}(\theta_0) - E\left\{\nabla L_{F,j}(\theta_0)\right\}\|_2 \leq \frac{\mu_-^2}{32MK^3}\right\},\\
\varepsilon_{F,3} &= \left\{\|E\left\{\nabla T(\theta_0)\right\}\|_2 \leq \frac{\mu_-}{4}\right\}.
\end{aligned}$$

For any $\theta \in \mathcal{A}$, where $\mathcal{A} = \left\{\|\theta - \theta_0\|_2 < \mu_-(4C_1MK^3)^{-1}\right\}$ where C_1 is a constant, we apply the Taylor's Expansion, which gives us

$$\begin{aligned}
L_{F,1}(\theta) - L_{F,1}(\theta_0) - \langle\nabla L_{F,1}(\theta_0), \theta - \theta_0\rangle &= (\theta-\theta_0)^{\mathrm{T}}\int_0^1 \nabla^2 L_{F,1}\left\{\theta + v(\theta_0 - \theta)\right\}dv\\
&\geq \lambda_{\min}[\int_0^1 \nabla^2 L_{F,1}\left\{\theta + v(\theta_0 - \theta)\right\}dv]\|\theta - \theta_0\|_2^2
\end{aligned}$$

Then, we look at the first item

$$
\begin{aligned}
& \lambda_{\min}[\int_0^1 \nabla^2 L_{F,1}\left\{\theta + v(\theta_0 - \theta)\right\} dv] \\
\geq\ & \lambda_{\min}[E\left\{\nabla^2 L_1(\theta_0)\right\}] - \| \int_0^1 \nabla^2 L_{F,1}\left\{\theta + v(\theta_0 - \theta)\right\} dv - E\left\{\nabla^2 L_1(\theta_0)\right\} \|_2 \\
\geq\ & \mu_- - \| \int_0^1 \nabla^2 L_{F,1}\left\{\theta + v(\theta_0 - \theta)\right\} dv - \nabla^2 L_{F,1}(\theta_0)\|_2 - \|\nabla^2 L_{F,1}(\theta_0) - E\left\{\nabla^2 L_{F,1}(\theta_0)\right\} \|_2 \\
& -\|E\left\{\nabla^2 L_{F,1}(\theta_0)\right\} - E\left\{\nabla^2 L_1(\theta_0)\right\} \|_2 \\
\geq\ & \mu_- - \| \int_0^1 \nabla^2 L_{F,1}\left\{\theta + v(\theta_0 - \theta)\right\} dv - \nabla^2 L_{F,1}(\theta_0)\|_2 - \|\nabla^2 L_{F,1}(\theta_0) - E\left\{\nabla^2 L_{F,1}(\theta_0)\right\} \|_2 \\
& -\|E\left\{\nabla T(\theta_0)\right\} \|_2 \\
=\ & \mu_- - \frac{\mu_-}{4} - \frac{\mu_-}{4} - \frac{\mu_-}{4} \\
=\ & \frac{\mu_-}{4}
\end{aligned}
$$

The first inequality is by $\lambda_{\min}(A) \geq \lambda_{\min}(B) - \|A - B\|_2$ for two any symmetric and semi-definite matrices A and B. The last step results from the application of Lemma B.3.6. Thus, by choosing $\theta = \hat{\theta}_{F,1} + v(\theta_0 - \hat{\theta}_{F,1})$ we have under event $\varepsilon = \varepsilon_{F,0} \cap \varepsilon_{F,1} \cap \varepsilon_{F,2}$, we have

$$\|\hat{\theta}_{F,1} - \theta_0\|_2 \leq 4\mu_-^{-1} \|\nabla L_{F,1}(\theta_0) - E\left\{\nabla L_{F,1}(\theta_0)\right\} \|_2.$$

From Lemma B.3.15, we have that

$$\|\nabla L_{F,1}(\theta_0) - E\left\{\nabla L_{F,1}(\theta_0)\right\} \|_2 \leq C_1 \left\{(p+1) \vee \log \alpha_n\right\}^{1/2} \alpha_n^{-1/2}.$$

This finishes the proof.

□

Lemma B.3.9. *Under assumptions 1 to 3, for any $\xi > 0$ and any $\theta \in \mathbf{R}^{p+1}$, we have with probability*

at least $1 - 4m\alpha_n^{-8} - 2mn^{-8}$ *that*

$$\|\hat{\theta} - \theta_0\|_2 \leq C\mu_-^{-1}(\phi M)^{1/2} K \left\{ \frac{(p+1) \vee 8 \log \alpha_n}{m\alpha_n} \right\}^{1/2}$$

Proof. We begin to define the following events for each site,

$$\begin{aligned}
\varepsilon_j = & \left\{ \max_{\|u\|_2=1} \frac{1}{n} \sum_{i=1}^{n} (X_{ij}^{\mathrm{T}} u)^4 \leq C_1 K^4 \right\} \cap \\
& \left\{ \|\nabla^2 L_{F,j}(\theta_0) - E\left\{\nabla^2 L_{F,j}(\theta_0)\right\}\|_2 \leq \frac{\mu_-}{4} \right\} \cap \\
& \left\{ \|\nabla L_{F,j}(\theta_0) - E\left\{\nabla L_{F,j}(\theta_0)\right\}\|_2 \leq \frac{\mu_-^2 K}{32MK^3} \right\}.
\end{aligned}$$

By applying the Lemma 6 from Zhang et al. (2012), we have

$$\|\hat{\theta} - \theta_0\|_2 \leq 2\mu_-^{-1} \|\nabla L(\theta_0) - E\left\{\nabla L(\theta_0)\right\}\|_2.$$

By applying the Lemma B.3.1 with choice of $\xi = 8 \log \alpha_n$, we have

$$\|\hat{\theta} - \theta_0\|_2 \leq C_1\mu_-^{-1}(\phi M)^{1/2} K \left\{ \frac{(p+1) \vee 8 \log \alpha_n}{m\alpha_n} \right\}^{1/2}.$$

This completes the proof. □

Lemma B.3.10. *Under assumption 1, for* j*th site, with probability* $1 - 2\exp(\xi)$*, we have*

$$\max_{\|u\|_2^2=1} \left\{ n^{-1} \sum_{i=1}^{n} (X_{ij}^{\mathrm{T}} u)^4 \right\} \lesssim K^4.$$

Proof. By the triangle inequality, we have

$$\max_{\|u\|_2=1}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}u)^4\right\} \leq \max_{\|u\|_2=1}[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_{ij}^{\top}u)^4 - E\left\{\tfrac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}u)^4\right\}] + \max_{\|u\|_2=1} E\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_{ij}^{\top}u)^4\right\}.$$

Given any u, by sub-Gaussianity, we have $E\exp[\left\{(X_{ij}^{\top}u)^4K^4\right\}^{1/2}] \leq 2$, from which we deduce that $\|(X_{ij}^{\top}u)^4\|_{\psi_{1/2}} \leq K^4$ and $E\left\{(X_{ij}^{\top}u)^4\right\} \leq C_1K^4$ where C_1 is a constant. We can write $u = v + \delta$ for some v in $N(1/16)$ and an error vector δ such that $\|\delta\|_2 \leq 1/16$.

$$\max_{\|u\|_2=1}[\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}u)^4 - E\left\{\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}u)^4\right\}] \leq 2\max_{v\in N(1/16)}[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_{ij}^{\top}v)^4 - E\left\{\tfrac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}v)^4\right\}].$$

Applying the bound above display (3.6) in Adamczak et al. (2011), we have for any $t > 0$ that

$$P[|\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}v)^4 - E\left\{(X_{ij}^{\top}v)^4\right\}| \geq C_2K^4\left\{(\frac{t}{n})^{1/2} + \frac{t^2}{n}\right\}] \leq 2e^{-(t-3)}, \tag{B.1}$$

where C_2 is a constant. Applying a union bound over $v \in N(1/16)$, we deduce that

$$P(\max_{v\in N^d(1/16)}[\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}v)^4 - E\left\{\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}v)^4\right\}] \geq C_3 K^4\left\{(\tfrac{t}{n})^{1/2} + \tfrac{t^2}{n}\right\}) \leq 2e^{-(t-3-p\log 33)},$$

where C_3 is a constant. Substituting $\xi = t - 3 - p\log 33$ into it with positive ξ and combining the equations, we derive that with probability at least $1 - 2e^{-\xi}$ such that

$$\max_{\|u\|_2=1}[\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}u)^4 - E\left\{\frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{\top}u)^4\right\}] \lesssim K^4\left\{(\tfrac{p\vee\xi}{n})^{1/2} + \tfrac{(p\vee\xi)^2}{n}\right\}.$$

Let $\xi = 2\log n$. For a sufficiently large n satisfied that $n \geq \max(p^2, \log^2 n)$, we have with probability

at least $1 - 2n^{-2}$ that

$$\max_{\|u\|_2=1} \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{ij}^\top u)^4 \right\} \lesssim K^4.$$

□

Lemma B.3.11. *Under Assumptions 1-3 and event ε, the mean square error of B_4 has the following bound,*

$$E(\|B_4\|_2^2) \leq C \frac{\{(p+1) \vee 4 \log \alpha_n\}^3}{\alpha_n^3},$$

where

$$\begin{aligned} C &= 6\mu_-^{-6}\phi M^5 K^{10} + 6\mu_-^{-2}\|\Theta^{-1}\|_2^2 \phi^2 M^4 K^8 + (3/2)\mu_-^{-2}\|\Theta^{-1}\|_2^2 \phi^2 M^4 K^8 \\ &\quad + (3/2)\|\Theta^{-2}\|_2^2 \mu_+^2 \phi^3 M^3 K^6 + (3/2)\mu_-^{-6}\phi^2 M^4 K^{10} + (1/6)\|\Upsilon^{-1}\|_2^2 \phi^3 M^3 K^6. \end{aligned}$$

Proof. We used the index notation from McCullagh (2018) to define the components of B_4. Specially, we used the G_r to denote the rth element of $\nabla L_1(\theta)$, $G_{r,s}$ for the (r,s)th element of the matrix $\nabla^2 L_1(\theta) - \Sigma$, $G_{r,s,t}$ to denote the (r,s,t) element of the tensor $\nabla^3 L_1 - \Theta$, and $G_{r,s,t,u}$ to denote the (r,s,t,u) element of the tensor $R_4 - \Upsilon$. We use the the Equation (7.10) in McCullagh (2018), which gives us the rth element of B_4,

$$\begin{aligned} (B_4)_r &= \underbrace{I_{r,s}^{-1} I_{t,u}^{-1} I_{v,w}^{-1} G_{s,t} G_{u,v} G_w}_{(B_{41})_r} + \underbrace{\Theta_{r,s,t}^{-1} I_{u,v}^{-1} G_s G_{t,u} G_v}_{(B_{42})_r} + \underbrace{(1/2) I_{r,s}^{-1} \Theta_{t,u,v}^{-1} G_{s,t} G_u G_v}_{(B_{43})_r} \\ &\quad + \underbrace{(1/2)\Theta_{r,s,t}^{-1} \Theta_{u,v,w}^{-1} I_{t,w} G_s G_u G_w}_{(B_{44})_r} + \underbrace{(1/2) I_{r,s}^{-1} I_{t,u}^{-1} I_{v,w}^{-1} G_{s,u,w} G_t G_v}_{(B_{45})_r} \\ &\quad + \underbrace{(1/6)\Upsilon_{r,s,t,u}^{-1} G_s G_t G_u}_{(B_{46})_r}. \end{aligned} \tag{B.2}$$

Then we can have

$$E\left\{\|B_4\|_2^2\right\} \leq \sum_{t=1}^{6} 6E\left\{\|B_{4t}\|_2^2\right\}. \tag{B.3}$$

We then investigate B_{4t} for $t = 1, ..., 6$.

We first show the bound of B_{41},

$$\begin{aligned} E(\|B_{41}\|_2^2) &\leq \|I^{-3}\|_2^2([E\left\{\|(\nabla^2 L_1(\theta_0)-\Sigma)\|_2^4\right\}]^2 E\left\{\|\nabla L_1(\theta_0)\|_2^4\right\})^{1/2} \\ &\leq \mu_-^{-6}\phi M^5 K^{10}(p\vee\log\alpha_n)^3\alpha_n^{-3}. \end{aligned} \tag{B.4}$$

As for B_{42}, we have

$$\begin{aligned} E(\|B_{42}\|_2^2) &\leq \|I^{-1}\|_2^2\|\Theta^{-1}\|_2^2(E\left\{\|(\nabla^2 L_1(\theta_0)-\Sigma)\|_2^4\right\}[E\left\{\|\nabla L_1(\theta_0)\|_2^4\right\}]^2)^{1/2} \\ &\leq \mu_-^{-2}\|\Theta^{-1}\|_2^2\phi^2 M^4 K^8(p\vee\log\alpha_n)^3\alpha_n^{-3}. \end{aligned} \tag{B.5}$$

Similarly, for B_{43}, we have

$$\begin{aligned} E(\|B_{43}\|_2^2) &\leq (1/4)\|I^{-1}\|_2^2\|\Theta^{-1}\|_2^2(E\left\{\|(\nabla^2 L_1(\theta_0)-\Sigma)\|_2^4\right\}[E\left\{\|\nabla L_1(\theta_0)\|_2^4\right\}]^2)^{1/2} \\ &\leq (1/4)\mu_-^{-2}\|\Theta^{-1}\|_2^2\phi^2 M^4 K^8(p\vee\log\alpha_n)^3\alpha_n^{-3}. \end{aligned} \tag{B.6}$$

We apply the Lemma B.3.1, Lemma B.3.2, Lemma B.3.3, and Lemma B.3.4 to B_{44}, which gives us

$$\begin{aligned} E(\|B_{44}\|_2^2) &\leq (1/4)\|\Theta^{-2}\|_2^2\|I\|_2^2([E\left\{\|\nabla L_1(\theta_0)\|_2^4\right\}]^3)^{1/2} \\ &\leq (1/4)\|\Theta^{-2}\|_2^2\mu_+^2\phi^3 M^3 K^6(p\vee\log\alpha_n)^3\alpha_n^{-3}. \end{aligned} \tag{B.7}$$

Similarly for B_{45} with Lemma B.3.2 and B.3.3, we have

$$\begin{aligned} E(\|B_{45}\|_2^2) &\leq (1/4)\|I^{-3}\|_2^2([E\left\{\|\nabla L_1(\theta_0)\|_2^4\right\}]^2 E\|\nabla^3 L_1(\theta_0)-\Theta\|_2^4)^{1/2} \\ &\leq (1/4)\mu_-^{-6}\phi^2 M^4 K^{10}(p\vee\log\alpha_n)^3\alpha_n^{-3}. \end{aligned} \tag{B.8}$$

By applying Lemma B.3.1 to B_{46}, which gives us

$$\begin{aligned} E(\|B_{46}\|_2^2) &\leq (1/36)\|\Upsilon^{-1}\|_2^2([E\left\{\|\nabla L_1(\theta_0)\|_2^4\right\}]^3)^{1/2} \\ &\leq (1/36)\|\Upsilon^{-1}\|_2^2\phi^3 M^3 K^6 (p \vee \log\alpha_n)^3 \alpha_n^{-3}. \end{aligned} \tag{B.9}$$

Combining the results from Equation (B.4), Equation (B.5), Equation (B.6), Equation (B.7), Equation (B.8), and Equation (B.9) by Equation (B.3), we have

$$E\left\{\|B_4\|_2^2\right\} \leq C\left\{(p+1) \vee 4\log\alpha_n\right\}^3 \alpha_n^{-3}, \tag{B.10}$$

where

$$\begin{aligned} C &= 6\mu_-^{-6}\phi M^5 K^{10} + 6\|I^{-1}\|_2^2\|\Theta^{-1}\|_2^2\phi^2 M^4 K^8 + (3/2)\mu_-^{-2}\|\Theta^{-1}\|_2^2\phi^2 M^4 K^8 \\ &\quad +(3/2)\|\Theta^{-2}\|_2^2\mu_+^2\phi^3 M^3 K^6 + (3/2)\mu_-^{-6}\phi^2 M^4 K^{10} + (1/6)\|\Upsilon^{-1}\|_2^2\phi^3 M^3 K^6. \end{aligned}$$

This suggests that the bound for the B_{41} to B_{46} has the same order, and thus, we can rewrite the bound for B_4 as a constant times $\left\{(p+1) \vee 4\log\alpha_n\right\}^3 \alpha_n^{-3}$. □

Lemma B.3.12. *Under Assumptions 1-3 and the event ε, the bias of B_4 has the following bound,*

$$\|E(B_4)\|_2^2 \leq C\frac{(p+1)^4}{\alpha_n^4},$$

where $C = 6\mu_-^{-6}\phi M^5 K^{10} + 6\|\Theta^{-1}\|_2^2\mu_-^2\phi^2 M^4 K^8 + (3/2)\alpha_n^{-4}\mu_-^{-4}\phi^2 M^4 K^8 + (3/2)\mu_+^2\mu_-^{-4} M^2 K^6 + (3/2)\mu_-^{-6} M^4 K^{10} + (1/6)\|\Upsilon_{r,s,t,u}^{-1}\|_2^2 M^2 K^6$.

Proof. To prove the Lemma B.3.12, we can use the Equation (B.2), and thus we have

$$\|E(B_4)\|_2^2 \leq \sum_{t=1}^{t} 6\|E(B_{4t})\|_2^2. \tag{B.11}$$

We will look at the bound for B_{4t}, $t = 1, ..., 6$. To start, we first define the functions $U_s(X, Y) = \nabla_{\theta_s} L(\theta_0; X)$, with $\nabla^e_{\theta_s} g(\theta)$ indicating the eth order derivative of a function g with respect to the θ_s. We also define $U_{s,t}(X) = g_2(Z; \beta_0) X_s X_t - \Sigma_{s,t}$ and $U_{s,t,u}(X) = g_2(Z; \beta_0) X_s X_t X_u - \Theta_{s,t,u}$. These functions are mapping the space of vector X to the $\mathbb{R}^{(p+1)\times(p+1)}$ and $\mathbb{R}^{(p+1)\times(p+1)\times(p+1)}$.

$$\begin{aligned} E\left\{(B_{41})_r\right\} &= E(I^{-1}_{r,s} I^{-1}_{t,u} I^{-1}_{v,w} G_{s,t} G_{u,v} G_w) \\ &= \alpha_n^{-3} E\left\{ I^{-1}_{r,s} I^{-1}_{t,u} I^{-1}_{v,w} \sum_{i=1}^{n} U_{s,t}(X_{i1}) U_{u,v}(X_{i1}) U_w(X_{i1}, Y_{i1}) \right\} \\ &= \alpha_n^{-2} E\left\{ I^{-1}_{r,s} I^{-1}_{t,u} I^{-1}_{v,w} U_{s,t}(X_{i1}) U_{u,v}(X_{i1}) U_w(X_{i1}, Y_{i1}) \right\} \end{aligned}$$

Then, by Jensen's inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned} \|E\left\{(B_{41})_r\right\}\|_2^2 &= \alpha_n^{-4} \|E\left\{ I^{-1}_{r,s} I^{-1}_{t,u} I^{-1}_{v,w} U_{s,t}(X_{i1}) U_{u,v}(X_{i1}) U_w(X_{i1}, Y_{i1}) \right\}\|_2^2 \\ &\leq \alpha_n^{-4} [E\left\{ \|I^{-1}_{r,s} I^{-1}_{t,u} I^{-1}_{v,w} U_{s,t}(X_{i1}) U_{u,v}(X_{i1}) U_w(X_{i1}, Y_{i1})\|_2 \right\}]^2 \\ &\leq \alpha_n^{-4} \|I^{-1}_{r,s} I^{-1}_{t,u} I^{-1}_{v,w}\|_2^2 E\left\{\|U_{s,t}(X_{i1})\|_2^2\right\} E\left\{\|U_{u,v}(X_{i1})\|_2^2\right\} \\ &\quad E\left\{\|U_w(X_{i1}, Y_{i1})\|_2^2\right\} \end{aligned}$$

This will give us

$$\|E(B_{41})\|_2^2 \leq \alpha_n^{-4} \mu_-^{-6} \phi M^5 K^{10} (p+1)^4 \tag{B.12}$$

We now examine the bound of the bias for B_{42}, which is

$$\begin{aligned} E\left\{(B_{42})_r\right\} &= E(\Theta^{-1}_{r,s,t} I^{-1}_{u,v} G_s G_{t,u} G_v) \\ &= \alpha_n^{-3} E\left\{ \Theta^{-1}_{r,s,t} I^{-1}_{u,v} \sum_{i=1}^{n} U_s(X_{i1}, Y_{i1}) U_{t,u}(X_{i1}) U_v(X_{i1}, Y_{i1}) \right\} \\ &= \alpha_n^{-2} E\left\{ \Theta^{-1}_{r,s,t} I^{-1}_{u,v} U_s(X_{i1}, Y_{i1}) U_{t,u}(X_{i1}) U_v(X_{i1}, Y_{i1}) \right\} \end{aligned}$$

We then apply the Jensen's inequality and Cauchy-Schwarz inequality, which gives us

$$\begin{aligned}
\|E\left\{(B_{42})_r\right\}\|_2^2 &= \alpha_n^{-4}\|E\left\{\Theta_{r,s,t}^{-1}I_{u,v}^{-1}U_s(X_{i1},Y_{i1})U_{t,u}(X_{i1})U_v(X_{i1},Y_{i1})\right\}\|_2^2 \\
&\leq \alpha_n^{-4}[E\left\{\|\Theta_{r,s,t}^{-1}I_{u,v}^{-1}U_s(X_{i1},Y_{i1})U_{t,u}(X_{i1})U_v(X_{i1},Y_{i1})\|_2\right\}]^2 \\
&\leq \alpha_n^{-4}\|\Theta_{r,s,t}^{-1}\|_2^2\|I_{u,v}^{-1}\|_2^2 E\left\{\|U_s(X_{i1},Y_{i1})\|_2^2\right\}E\left\{\|U_{t,u}(X_{i1})\|_2^2\right\} \\
&\quad E\left\{\|U_v(X_{i1},Y_{i1})\|_2^2\right\}.
\end{aligned}$$

This will give us

$$\|E(B_{42})\|_2^2 \leq \alpha_n^{-4}\|\Theta^{-1}\|_2^2\mu_-^{-2}\phi^2M^4K^8p^4. \tag{B.13}$$

We continue our investigation for B_{43}. Notice that

$$\begin{aligned}
E\left\{(B_{43})_r\right\} &= E\left\{(1/2)I_{r,s}^{-1}\Theta_{t,u,v}^{-1}G_{s,t}G_uG_v\right\} \\
&= (1/2)\alpha_n^{-3}E\left\{I_{r,s}^{-1}\Theta_{t,u,v}^{-1}\sum_{i=1}^{n}U_{s,t}(X_{i1})U_u(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\} \\
&= (1/2)\alpha_n^{-2}E\left\{I_{r,s}^{-1}\Theta_{t,u,v}^{-1}U_{s,t}(X_{i1})U_u(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\}.
\end{aligned}$$

And we have

$$\begin{aligned}
\|E\left\{(B_{43})_r\right\}\|_2^2 &= (1/4)\alpha_n^{-4}\|E\left\{I_{r,s}^{-1}\Theta_{t,u,v}^{-1}U_{s,t}(X_{i1})U_u(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\}\|_2^2 \\
&\leq (1/4)\alpha_n^{-4}[E\left\{\|I_{r,s}^{-1}\Theta_{t,u,v}^{-1}U_{s,t}(X_{i1})U_u(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\|_2\right\}]^2 \\
&\leq (1/4)\alpha_n^{-4}\|I_{r,s}^{-1}\|_2^2\|\Theta_{t,u,v}^{-1}\|_2^2 E\left\{\|U_{s,t}(X_{i1})\|_2^2\right\}E\left\{\|U_u(X_{i1},Y_{i1})\|_2^2\right\} \\
&\quad E\left\{\|U_v(X_{i1},Y_{i1})\|_2^2\right\}.
\end{aligned}$$

The first inequality can be acquired by Jensen's inequality and the second inequality can be achieved via the Cauchy-Schwartz inequality.

And this gives us

$$\|E(B_{43})\|_2^2 \leq (1/4)\alpha_n^{-4}\mu_-^{-2}\|\Theta^{-1}\|_2^2\phi^2 M^4 K^8 p^4. \tag{B.14}$$

Note that for B_{44}, we show the bias of it, which is

$$\begin{aligned}
E\left\{(B_{44})_r\right\} &= E\left\{(1/2)\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}I_{t,w}G_sG_uG_w\right\} \\
&= (1/2)\alpha_n^{-3}E\left\{\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}I_{t,w}\sum_{i=1}^{n}U_s(X_{i1},Y_{i1})U_u(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\} \\
&= (1/2)\alpha_n^{-2}E\left\{\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}I_{t,w}U_s(X_{i1},Y_{i1})U_u(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\} \\
&= (1/2)\alpha_n^{-2}\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}I_{t,w}E\left\{U_s(X_{i1},Y_{i1})U_u(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\} \\
&= -(1/2)\alpha_n^{-2}\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}I_{t,w}E\left\{U_{s,u,v}(X_{i1})\right\}.
\end{aligned}$$

The last equality is found by the third Bartlett identity.

We take the norm of the bias, and give us

$$\begin{aligned}
\|E\left\{(B_{44})_r\right\}\|_2^2 &= (1/4)\alpha_n^{-4}\|E\left\{\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}I_{t,w}U_{s,u,v}(X_{i1})\right\}\|_2^2 \\
&\leq (1/4)\alpha_n^{-4}[E\left\{\|\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}I_{t,w}U_{s,u,v}(X_{i1})\|_2\right\}]^2 \\
&\leq (1/4)\alpha_n^{-4}\|\Theta^{-1}_{r,s,t}\Theta^{-1}_{u,v,w}\|_2^2\|I_{t,w}\|_2^2E\left\{\|U_{s,u,v}(X_{i1})\|_2^2\right\}
\end{aligned}$$

The first inequality can be acquired by Jensen's inequality and the second inequality can be achieved via the Cauchy-Schwartz inequality.

And this gives us

$$\|E(B_{44})\|_2^2 \leq (1/4)\alpha_n^{-4}\mu_+^2\|\Theta^{-2}\|_2^2 M^2 K^6 p^4. \tag{B.15}$$

Similarly, we show the bound of the bias for B_{45},

$$\begin{aligned} E\left\{(B_{45})_r\right\} &= E\left\{(1/2)I_{r,s}^{-1}I_{t,u}^{-1}I_{v,w}^{-1}G_{s,u,w}G_tG_v\right\} \\ &= (1/2)\alpha_n^{-3}E\left\{I_{r,s}^{-1}I_{t,u}^{-1}I_{v,w}^{-1}\sum_{i=1}^{n}U_{s,u,w}(X_{i1})U_t(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\} \\ &= (1/2)\alpha_n^{-2}E\left\{I_{r,s}^{-1}I_{t,u}^{-1}I_{v,w}^{-1}U_{s,u,w}(X_{i1})U_t(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\} \end{aligned}$$

We then get bound by taking the norm of the bias for B_{45}, and give us

$$\begin{aligned} \|E\left\{(B_{45})_r\right\}\|_2^2 &= (1/4)\alpha_n^{-4}\|E\left\{I_{r,s}^{-1}I_{t,u}^{-1}I_{v,w}^{-1}U_{s,u,w}(X_{i1})U_t(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\}\|_2^2 \\ &\leq (1/4)\alpha_n^{-4}[E\left\{\|I_{r,s}^{-1}I_{t,u}^{-1}I_{v,w}^{-1}E\left\{U_{s,u,w}(X_{i1})U_t(X_{i1},Y_{i1})U_v(X_{i1},Y_{i1})\right\}\|_2\right\}]^2 \\ &\leq (1/4)\alpha_n^{-4}\|I_{r,s}^{-1}I_{t,u}^{-1}I_{v,w}^{-1}\|_2^2E\left\{\|U_{s,u,w}(X_{i1})\|_2^2\right\}E\left\{\|U_t(X_{i1},Y_{i1})\|_2^2\right\} \\ &\quad E\left\{\|U_v(X_{i1},Y_{i1})\|_2^2\right\} \end{aligned}$$

The first inequality and the second inequality can be found by Jensen's inequality and Cauchy-Schwartz inequality.

This implies that

$$\|E(B_{45})\|_2^2 \leq (1/4)\alpha_n^{-4}\mu_-^{-6}M^4K^{10}p^4. \tag{B.16}$$

Finally, for B_{46}, we show the bias of it, which displays as

$$\begin{aligned} E\left\{(B_{46})_r\right\} &= E\left\{(1/6)\Upsilon_{r,s,t,u}^{-1}G_sG_tG_u\right\} \\ &= (1/6)\alpha_n^{-3}E\left\{\Upsilon_{r,s,t,u}^{-1}\sum_{i=1}^{n}U_s(X_{i1},Y_{i1})U_t(X_{i1},Y_{i1})U_u(X_{i1},Y_{i1})\right\} \\ &= (1/6)\alpha_n^{-2}E\left\{\Upsilon_{r,s,t,u}^{-1}U_s(X_{i1},Y_{i1})U_t(X_{i1},Y_{i1})U_u(X_{i1},Y_{i1})\right\} \\ &= (1/6)\alpha_n^{-2}\Upsilon_{r,s,t,u}^{-1}E\left\{U_s(X_{i1},Y_{i1})U_t(X_{i1},Y_{i1})U_u(X_{i1},Y_{i1})\right\} \\ &= -(1/6)\alpha_n^{-2}\Upsilon_{r,s,t,u}^{-1}E\left\{U_{s,t,u}(X_{i1})\right\}. \end{aligned}$$

The last equality can be achieved by the third Bartlett identity.

We take the norm of the bias, and give us

$$\begin{aligned}\|E\left\{(B_{46})_r\right\}\|_2^2 &= (1/36)\alpha_n^{-4}\|E\left\{\Upsilon_{r,s,t,u}^{-1}U_{s,t,u}(X_{i1})\right\}\|_2^2 \\ &\leq (1/36)\alpha_n^{-4}[E\left\{\|\Upsilon_{r,s,t,u}^{-1}E\left\{U_{s,t,u}(X_{i1})\right\}\|_2\right\}]^2 \\ &\leq (1/36)\alpha_n^{-4}\|\Upsilon_{r,s,t,u}^{-1}\|_2^2 E\left\{\|U_{s,t,u}(X_{i1})\|_2^2\right\}\end{aligned}$$

The first inequality can be acquired by Jensen's inequality and the second inequality can be achieved via the Cauchy-Schwartz inequality.

And this gives us

$$\|E(B_{46})\|_2^2 \leq (1/36)\alpha_n^{-4}\|\Upsilon_{r,s,t,u}^{-1}\|_2^2 M^2K^6p^4. \tag{B.17}$$

Combining the Equation (B.12), Equation (B.13), Equation (B.14), Equation (B.15, Equation (B.16), and Equation (B.17), we have

$$\|E(B_4)\|_2^2 \leq C\left\{\frac{(p+1)\vee 4\log(\alpha_n)}{\alpha_n}\right\}^4,$$

where $C = 6\mu_-^{-6}\phi M^5K^{10} + 6\|\Theta^{-1}\|_2^2\mu_-^{-2}\phi^2M^4K^8 + (3/2)\alpha_n^{-4}\mu_-^{-4}\phi^2M^4K^8 + (3/2)\mu_+^2\mu_-^{-4}M^2K^6 + (3/2)\mu_-^{-6}M^4K^{10} + (1/6)\|\Upsilon_{r,s,t,u}^{-1}\|_2^2M^2K^6$.

□

Lemma B.3.13. *Under Assumptions 1 and 3, we have that*

$$\|E(B_2+B_3+B_5)\|_2^2 \leq \|I^{-3}\|_2^2M^4K^{10}(p\vee\alpha_n)^4\alpha_n^{-4}$$

where $B_2 = I^{-2}(\nabla^2L_1(\theta_0)-\Sigma)\nabla L_1(\theta_0)$, $B_3 = 1/2\Theta^{-1}\nabla L_1(\theta_0)\otimes\nabla L_1(\theta_0)$, $B_5 = I^{-1}\alpha_n^{-1}b(\theta_0)$, C_2 *and* C_3 *are constants.*

Proof. We use the index notation and define

$$(B_{51})_r = I^{-1}_{r,s} I^{-1}_{t,u} \nabla_{\theta_u} I_{s,t}(\theta_0)$$

and

$$(B_{52})_r = I^{-1}_{r,s} [\alpha_n^{-1} \left\{ I^{-1}_{t,u}(\theta_0) - \alpha_n I^{-1}_{t,u} \right\} \nabla_{\theta_u} I_{s,t}(\theta_0)]$$

. We can decompose $(B_5)_r$ into

$$(B_5)_r = (B_{51})_r + (B_{52})_r.$$

We first show that $E\left\{(B_2)_r + (B_3)_r + (B_{51})_r\right\} = 0$ and then we show that $E\left\{(B_{52})_r\right\}$ is as large as $O(\alpha_n^2)$.

The first part of proof follows the proof from Firth (1993). We start to write down the B_2, B_3, and B_{51} in the index notation with G_r, $G_{r,s}$, and $G_{r,s,t}$ defined in the Lemma B.3.11, which gives us,

$$\begin{aligned}
(B_2)_r &= I^{-1}_{r,s} I^{-1}_{t,u} G_{s,t} G_u, \\
(B_3)_r &= (1/2) \Theta^{-1}_{r,s,t} G_{s,t}, \\
(B_{51})_r &= -(1/2) I^{-1}_{r,s} I^{-1}_{t,u} (G_{s,t,u} + \Theta_{s,t,u}).
\end{aligned}$$

We first take the expectation for B_{51}, which gives us

$$E\left\{(B_{51})_r\right\} = -(1/2) I^{-1}_{r,s} I^{-1}_{t,u} \Theta_{s,t,u}. \tag{B.18}$$

We then take the expectation for $B_2 + B_3$, and we have

$$\begin{aligned} E\left\{(B_2)_r + (B_3)_r\right\} &= I_{r,s}^{-1} I_{t,u}^{-1} E(G_{s,t} G_u) + 1/2 \Theta_{r,s,t}^{-1} E(G_{s,t}) \\ &= -(1/2) I_{r,s}^{-1} I_{t,u}^{-1} \left\{E(G_s G_t G_u) + E(G_s G_{tu})\right\} \\ &= -(1/2) I_{r,s}^{-1} I_{t,u}^{-1} E(G_s G_t G_u) \\ &= (1/2) I_{r,s}^{-1} I_{t,u}^{-1} \Theta_{s,t,u} \end{aligned} \tag{B.19}$$

The second equality holds from the properties of maximum likelihood estimates shown in Page 210 of McCullagh (2018). The third equality holds by assumption 1-3 for the logistic regression model, $E(G_s G_{tu}) = 0$. The last equality holds because of the third Bartlett identity.

Combining Equation (B.18) and Equation (B.19), we get $E\left\{(B_2)_r + (B_3)_r + (B_{51})_r\right\} = 0$.

Then we define

$$U_{r,s,t}(X_{i1}) = \left\{I_{r,s}^{-1}(\theta_0) I_{t,u}^{-1}(\theta_0) (X_{i1})_r (X_{i1})_s - g_1(\alpha_0) I_{r,s}^{-1}\right\} (X_{i1})_s (X_{i1})_t (X_{i1})_u$$

and we find

$$\begin{aligned} \left\| E\left\{(B_{52})_r\right\} \right\|_2^2 &= \left\| \alpha_n^{-2} E[g_2^2(Z_{i1}; \beta_0) I_{r,s}^{-1} U_{s,t,u}(X_{i1})] \right\|_2^2 \\ &\leq \alpha_n^{-4} E \| g_2^2(Z_{i1}; \beta_0) I_{r,s}^{-1} U_{s,t,u}(X_{i1}) \|_2^2 \\ &\leq \mu_-^{-6} M^4 K^{10} p^4 \alpha_n^{-4}. \end{aligned}$$

This completes the proof. □

Lemma B.3.14. *Under Assumptions 1 to 3 and $n > (p+1)^2 \vee \log^2 n$, with the following events*

hold,

$$\begin{aligned}
\varepsilon_{F,0} &= \left\{ \max_{\|u\|_2=1} \frac{1}{n} \sum_{i=1}^{n} (X_{i1}^{\mathrm{T}} u)^4 \leq C_1 K^4 \right\}, \\
\varepsilon_{F,1} &= \left\{ \|\nabla^2 L_{F,j}(\theta_0) - E\left\{\nabla^2 L_{F,j}(\theta_0)\right\}\|_2 \leq \frac{\mu_-}{4} \right\}, \\
\varepsilon_{F,2} &= \left\{ \|\nabla L_{F,j}(\theta_0) - E\left\{\nabla L_{F,j}(\theta_0)\right\}\|_2 \leq \frac{\mu_-^2}{32MK^3} \right\}, \\
\varepsilon_{F,3} &= \left\{ \|E\left\{\nabla T(\theta_0)\right\}\|_2 \leq \frac{\mu_-}{4} \right\}.
\end{aligned}$$

we have the two-round distributed Firth and the pooled estimator satisfying that

$$\|\tilde{\theta}^{(2)} - \hat{\theta}_F\|_2 \leq 4\mu_-^{-1} \|\nabla \tilde{L}_F(\hat{\theta})\|_2$$

Proof. Under event $\varepsilon = \varepsilon_{F,0} \cap \varepsilon_{F,1} \cap \varepsilon_{F,2}$, we can apply the Lemma B.3.7 and we will have the results. This can be achieved by setting $\theta_0 = \hat{\theta}_F$ and substituting $L_{F,1}$ by $\tilde{L}_F$. We then need to check the following inequality holds,

$$\begin{aligned}
\|\nabla^2 L_{F,1}(\hat{\theta}_F) - E\left\{\nabla^2 L_1(\theta_0)\right\}\|_2^2 &\leq \frac{3\mu_-}{4}, \\
\|\nabla L_{F,1}(\hat{\theta}_F) - E\left\{\nabla L_{F,1}(\theta_0)\right\}\|_2^2 &\leq \frac{13\mu_-^2}{32MK^3}.
\end{aligned}$$

For any $\theta \in \mathcal{A}$, where $\mathcal{A} = \left\{\|\theta - \theta_0\|_2 < \mu_-(4C_1MK^3)^{-1}\right\}$, we start with the first inequality,

$$\begin{aligned}
\|\nabla^2 L_{F,1}(\hat{\theta}_F) - E\left\{\nabla^2 L_1(\theta_0)\right\}\|_2^2 &\leq \|\nabla^2 L_{F,1}(\hat{\theta}_F) - \nabla^2 L_{F,1}(\theta_0)\|_2 + \|\nabla^2 L_{F,1}(\theta_0) - E\left\{\nabla^2 L_{F,1}(\theta_0)\right\}\|_2 \\
&\quad + \|E\left\{\nabla^2 L_{F,1}(\theta_0)\right\} - E\left\{\nabla^2 L_1(\theta_0)\right\}\|_2 \\
&\leq \|\nabla^3 L_{F,1}(\hat{\theta}_F) - \nabla^3 L_{F,1}(\theta_0)\|_2 \|\hat{\theta}_F - \theta_0\|_2 + \frac{\mu_-}{4} + \frac{\mu_-}{4} \\
&= C_1 MK^3 \|\hat{\theta}_F - \theta_0\|_2 + \frac{\mu_-}{4} + \frac{\mu_-}{4} \\
&\leq \frac{3\mu_-}{4}.
\end{aligned}$$

The last equality is from the Lipschitz continuity of $\nabla^3 L_{F,1}(\theta)$ for any $\theta \in \mathcal{A}$ since it is a function of the first to the fourth order derivative of the function $L_1(\theta)$ and they are continuous as shown in

Lemma B.3.6.

For the second inequality, we apply the Taylor's expansion, which would give us,

$$\nabla L_{F,1}(\hat{\theta}_F) - \nabla L_{F,1}(\theta_0) = P_1(\hat{\theta}_F - \theta_0),$$

where matrix $P_1 = \int_0^1 \nabla^2 L_{F,1}\left\{\theta^* + v(\hat{\theta}_F - \theta_0)\right\} dv$ satistifes

$$\|P_1 - E\left\{\nabla^2 L_{F,1}(\theta_0)\right\}\|_2 \leq MK^3 \|\hat{\theta}_F - \theta_0\|_2.$$

By the triangle inequality, we deduce that

$$\begin{aligned}
\left\|\nabla L_{F,1}(\hat{\theta}_F) - \mathbb{E}\left\{\nabla L_{F,1}(\theta_0)\right\}\right\|_2 \leq & \left\|\nabla L_{F,1}(\theta_0) - \mathbb{E}\left\{\nabla L_{F,1}(\theta_0)\right\}\right\|_2 \\
& + \left\|P_1 - \mathbb{E}\left\{\nabla^2 L_{F,1}(\theta_0)\right\}\right\|_2 \cdot \left\|\hat{\theta}_F - \theta_0\right\|_2 \\
& + \left\|\mathbb{E}\left\{\nabla^2 L_{F,1}(\theta_0)\right\} - \mathbb{E}\left\{\nabla^2 L_1(\theta_0)\right\}\right\|_2 \cdot \left\|\hat{\theta}_F - \theta_0\right\|_2 \\
& + \left\|\mathbb{E}\left\{\nabla^2 L_1(\theta_0)\right\}\right\|_2 \cdot \left\|\hat{\theta}_F - \theta_0\right\|_2 \\
\leq & \frac{\mu_-^2}{32MK^3} + \frac{\mu_-^2}{16MK^3} + \frac{\mu_-^2}{16MK^3} + \frac{\mu_-^2}{4MK^3} \\
= & \frac{13\mu_-^2}{32MK^3}
\end{aligned}$$

Then, similar to the proof of B.3.8, we have,

$$\begin{aligned}
& \lambda_{\min}[\int_0^1 \nabla^2 \tilde{L}\left\{\hat{\theta}_F + v(\theta_0 - \hat{\theta}_F)\right\} dv] \\
= \; & \lambda_{\min}[\int_0^1 \nabla^2 L_{F,1}\left\{\hat{\theta}_F + v(\theta_0 - \hat{\theta}_F)\right\} dv] \\
\geq \; & \lambda_{\min}[E\left\{\nabla^2 L_1(\theta_0)\right\}] - \|\int_0^1 \nabla^2 L_{F,1}\left\{\hat{\theta}_F + v(\theta_0 - \hat{\theta}_F)\right\} dv - E\left\{\nabla^2 L_1(\theta_0)\right\}\|_2 \\
\geq \; & \mu_- - \frac{3\mu_-}{4} \\
= \; & \frac{\mu_-}{4}.
\end{aligned}$$

The first equality is by using $\nabla^2 \tilde{L}(\theta) = \nabla^2 L_{F,j}(\theta)$.

This shall give us

$$\|\tilde{\theta}^{(2)} - \hat{\theta}_F\|_2 \leq 4\mu_-^{-1} \|\nabla \tilde{L}_F(\hat{\theta}) - E\left\{\nabla \tilde{L}_F(\hat{\theta})\right\}\|_2 \leq 4\mu_-^{-1} \|\nabla \tilde{L}_F(\hat{\theta})\|_2$$

We complete this proof.

□

Lemma B.3.15. *Under Assumptions 1 to 3, $n > (p+1)^2 \vee \log^2 n$, and for any integer $k_0 \leq 16$, we have*

$$E[\|\nabla L_{F,j}(\theta_0) - E\left\{\nabla L_{F,j}(\theta_0)\right\}\|_2^{k_0}] \leq C\left\{(p+1) \vee \log \alpha_n\right\}^{k_0/2} \alpha_n^{-k_0/2}$$
$$E[\|\nabla^2 L_{F,j}(\theta_0) - E\left\{\nabla^2 L_{F,j}(\theta_0)\right\}\|_2^{k_0}] \leq C'\left\{(p+1) \vee \log \alpha_n\right\}^{k_0/2} \alpha_n^{-k_0/2},$$

where

$$\begin{aligned} C &= 2^{k_0}(\phi M)^{k_0/2} K^2 + 3(2^{k_0})(p+1)^{k_0/2+1} \mu_-^{k_0} M^{k_0/2} K^{3k_0/2} + 3(2^{k_0})(p+1)^{k_0/2+1} \\ &\quad \|\Theta^{-1}\|_2^{k_0/2} \mu_-^{2k_0} M^{k_0/2} K^{k_0} + 3(2^{k_0})(p+1)^{k_0/2+1} \mu_-^{2k_0} M^{k_0} K^{5k_0/2} \end{aligned}$$

and

$$\begin{aligned} C' &= M^{k_0} K^{k_0} + (2^{k_0})(p+1)^{k_0+3/2}(3\mu_-^{k_0} M^{k_0/2} K^{2k_0} \\ &\quad +3\|\Upsilon^{-1}\|_2^{k_0/2} \mu_-^{2k_0} M^{k_0/2} K^{k_0} + 3\mu_-^{2k_0} M^{k_0} K^{2k_0} \\ &\quad +5\mu_-^{3k_0} M^{k_0/2} K^{3k_0/2} + 5\|\Theta^{-1}\|_2^{k_0} \mu_-^{3k_0} M^{k_0/2} K^{k_0} \\ &\quad +5\|\Theta^{-1}\|_2^{k_0/2} \mu_-^{2k_0} M^{k_0/2} K^{3k_0/2} + 5\|\Theta^{-2}\|_2^{k_0/2} \mu_-^{3k_0} M^{k_0/2} K^{k_0} \\ &\quad +5\mu_-^{8k_0} M^{k_0} K^{3k_0}) \end{aligned}$$

Proof. We define the following events,

$$\begin{aligned}
\varepsilon_1 &= \left\{ \|\nabla L_1(\theta_0)\|_2 \leq (\phi M)^{1/2} K\Lambda \right\}, \\
\varepsilon_2 &= \left\{ \|\nabla^2 L_1(\theta_0) - E\left\{\nabla^2 L_1(\theta_0)\right\}\|_2 \leq MK^2\Lambda \right\}, \\
\varepsilon_3 &= \left\{ \|\nabla^3 L_1(\theta_0) - E\left\{\nabla^3 L_1(\theta_0)\right\}\|_2 \leq MK^3\Lambda \right\}, \\
\varepsilon_4 &= \left\{ \|\nabla^4 L_1(\theta_0) - E\left\{\nabla^4 L_1(\theta_0)\right\}\|_2 \leq MK^4\Lambda \right\}.
\end{aligned}$$

Under event $\varepsilon = \varepsilon_1 \cup \varepsilon_2 \cup \varepsilon_3 \cup \varepsilon_4$, we have

$$\begin{aligned}
E[\|\nabla L_{F,1}(\theta_0) - E\left\{\nabla L_{F,1}(\theta_0)\right\}\|_2^{k_0}] \;\leq\; & 2^{k_0} E\left\{ \|\nabla L(\theta_0)\|_2^{k_0} \right\} + \\
& 2^{k_0} E[\|T(\theta_0) - E\left\{T(\theta_0)\right\}\|_2^{k_0}] + P(\varepsilon^c) \qquad \text{(B.20)}
\end{aligned}$$

By Lemma B.3.1, we have $E\left\{ \|\nabla L(\theta_0)\|_2^{k_0} \right\} \leq C_1^{k_0} (\phi M)^{k_0/2} K^2 (p \vee \log_n)^{k_0/2} \alpha_n^{k_0/2}$. Now we look at the $E[\|T(\theta_0) - E\left\{T(\theta_0)\right\}\|_2^{k_0}]$. With the vector and matrix notation, we can define a vector D with the rth element $(r = 1, ..., p+1)$, $D_r = \alpha_n^{-1}(tr\left\{I(\theta_0)^{-1}\nabla_{\theta_r} I(\theta_0)\right\} - E[tr\left\{I(\theta_0)^{-1}\nabla_{\theta_r} I(\theta_0)\right\}])$

We can deduce that

$$\begin{aligned}
2^{k_0} E\left\{T(\theta_0)\right\}\|_2^{k_0} &= E(\|D\|_2^{k_0}) \\
&\leq E[\left\{ (p+1) \max_{1 \leq r \leq (p+1)} |D_r|^2 \right\}^{k_0/2}] \\
&\leq (p+1)^{k_0/2} \sum_{r=1}^{p+1} E(|D_r|^{k_0/2}).
\end{aligned}$$

We then can only look at $E(|D_r|^{k_0/2})$. Notice that

$$E(|D_r|^{k_0/2}) \leq \alpha_n^{-k_0} tr(E[\|I(\theta_0)^{-1}\nabla_{\theta_r} I(\theta_0) - E\left\{I(\theta_0)^{-1}\nabla_{\theta_r} I(\theta_0)\right\}\|_2^{k_0}]).$$

We denote $D_1 = \nabla_{\theta_r} I(\theta_0)$ and $D_2 = I^{-1}(\theta_0)$, and we have

$$\begin{aligned} E\left\{\|D_1 D_2 - E(D_1 D_2)\|_2^{k_0}\right\} &= E(\|\{D_1 - E(D_1)\} D_2 + E(D_1)\{D_2 - E(D_2)\} - \\ &\quad E\{D_1 D_2 - E(D_1)E(D_2)\}\|_2^{k_0}) \\ &\leq 3E[\|\underbrace{\{D_1 - E(D_1)\} D_2}_{F_1}\|_2^{k_0}] + 3E[\|\underbrace{E(D_1)\{D_2 - E(D_2)\}}_{F_2}\|_2^{k_0}] \\ &\quad +3\|E\left\{\underbrace{D_1 D_2 - E(D_1)E(D_2)}_{F_3}\right\}\|_2^{k_0}. \end{aligned}$$

For the item F_1, by Jensen's inequality and Holder's inequality, we have

$$\begin{aligned} E(\|F_1\|_2^2) &\leq [E\{\|D_2\|_2\}]^{k_0/2}[E\{\|D_1 - E(D_1)\|_2\}]^{k_0/2} \\ &\leq \mu_-^{k_0} M^{k_0/2} K^{3k_0/2}\{(p+1) \vee \log \alpha_n\}^{k_0/2} \alpha_n^{-k_0/2} \end{aligned} \tag{B.21}$$

Similarly, for F_2, we have

$$\begin{aligned} E(\|F_2\|_2^2) &\leq [E\{\|E(D_1)\|_2\}]^{k_0/2}[E\{\|D_2 - E(D_2)\|_2\}]^{k_0/2} \\ &\leq \|\Theta^{-1}\|_2^{k_0/2} \mu_-^{2k_0} M^{k_0/2} K^{k_0/}\{(p+1) \vee \log \alpha_n\}^{k_0/2} \alpha_n^{-k_0/2}. \end{aligned} \tag{B.22}$$

For the bound of the bias for F_3, we define

$$U(X_{i1}) = g_2(\beta_0; Z_{i1}) X_{i1} X_{i1} X_{i1r} I^{-1}(\theta_0) g_2(\beta_0; Z_{i1}) X_{i1} X_{i1}^{\mathrm{T}} I^{-1}(\theta_0) - c_0,$$

where $c_0 = E(D_1)E(D_2)$ and we have

$$\begin{aligned} \|E(F_3)\|_2^{k_0}) &\leq \alpha_n^{-k_0/2} \|E\{U(X_{i1})\}\|_2^{k_0} \\ &\leq \alpha_n^{-k_0/2} E\|U(X_{i1})\|_2^{k_0} \\ &\leq \mu_-^{2k_0} M^{k_0} K^{5k_0/2}\{(p+1) \vee \log \alpha_n\}^{k_0/2} \alpha_n^{-k_0/2}. \end{aligned} \tag{B.23}$$

Also, when the event ε does not hold, we can reach out by setting $\xi = 8\log\alpha_n$ Lemma $B.3.1$, Lemma $B.3.2$, Lemma $B.3.3$, and Lemma $B.3.4$. Thus, when the order $k_0 \leq 16$, the bound holds.

Combining the results from Equation (B.21), Equation (B.22), and Equation (B.23) by Equation (B.20, we complete the first part of the proof.

For the bound for $\nabla^2 L_F(\theta_0)$, we start with the item $E[\|\nabla^2 L_F(\theta_0) - E\left\{\nabla^2 L_F(\theta_0)\right\}\|_2^{k_0}]$. Note that under event ε, we have

$$\begin{aligned} E[\|\nabla^2 L_F(\theta_0) - E\left\{\nabla^2 L_F(\theta_0)\right\}\|_2^{k_0}] &= 2^{k_0} E[\|\nabla^2 L(\theta_0) - E\left\{\nabla^2 L(\theta_0)\right\}\|_2^{k_0}] + \\ & \quad 2^{k_0} E[\|\nabla T(\theta_0) - E\left\{\nabla T(\theta_0)\right\}\|_2^{k_0}] \end{aligned} \tag{B.24}$$

The first part can be directly justified by Lemma B.3.2. We denote $D_{s,t}$ as the element for the sth row and tth column of the matrix $\nabla T(\theta_0)$. The second part can be continued with

$$\begin{aligned} E[\|\nabla T(\theta_0) - E\left\{\nabla T(\theta_0)\right\}\|_2^{k_0}] &\leq -E[\|\nabla T(\theta_0) - E\left\{\nabla T(\theta_0)\right\}\|_F^{k_0}] \\ &= E\left\{(\sum_{s=1}^{p+1}\sum_{t=1}^{p+1} D_{s,t}^2)^{k_0/2}\right\} \\ &\leq (p+1)^{k_0} E\left\{\max_{1\leq s\leq (p+1)}\max_{1\leq t\leq (p+1)} D_{s,t}^{k_0}\right\} \\ &\leq (p+1)^{k_0} E(\sum_{s=1}^{(p+1)}\sum_{t=1}^{(p+1)} |D_{s,t}|^{k_0}) \\ &= (p+1)^{k_0} \sum_{s=1}^{(p+1)}\sum_{t=1}^{(p+1)} E(|D_{s,t}|^{k_0}) \end{aligned}$$

We denote matrices in the following way,

$$G_{s,t} = I^{-1}(\theta_0)\nabla_{\theta_s}\nabla_{\theta_t} I(\theta_0)$$

and

$$H_{s,t} = \nabla_{\theta_s} I(\theta_0) I^{-1}(\theta_0)\nabla_{\theta_t} I(\theta_0) I^{-1}(\theta_0)$$

. Then, we have

$$
\begin{aligned}
(p+1)^{k_0}\sum_{s=1}^{p+1}\sum_{t=1}^{p+1}E(|D_{s,t}|^{k_0}) &= (p+1)^{k_0}\alpha_n^{-k_0}\sum_{s=1}^{p+1}\sum_{t=1}^{p+1}E[|tr(G_{s,t})-tr(H_{s,t}) \\
&\quad -E\left\{tr(G_{s,t})\right\}+E\left\{tr(H_{s,t})\right\}|^{k_0}] \\
&\leq (p+1)^{k_0}\alpha_n^{-k_0}\sum_{s=1}^{p+1}\sum_{t=1}^{p+1}E[tr\left\{|G_{s,t}-E(G_{s,t})|\right\}] \\
&\quad +E[tr\left\{|H_{s,t}-E(H_{s,t})|\right\}] \\
&\leq (p+1)^{k_0}\alpha_n^{-k_0}\sum_{s=1}^{p+1}\sum_{t=1}^{p+1}E\left\{\|G_{s,t}-E(G_{s,t})\|_1^{k_0}\right\} \\
&\quad +E\left\{\|H_{s,t}-E(H_{s,t})\|_1^{k_0}\right\} \\
&\leq (p+1)^{k_0+1}\alpha_n^{-k_0}\sum_{s=1}^{p+1}\sum_{t=1}^{p+1}E\left\{\|G_{s,t}-E(G_{s,t})\|_F^{k_0}\right\} \\
&\quad +E\left\{\|H_{s,t}-E(H_{s,t})\|_F^{k_0}\right\} \\
&\leq (p+1)^{k_0+3/2}\alpha_n^{-k_0}\sum_{s=1}^{p+1}\sum_{t=1}^{p+1}E\left\{\|G_{s,t}-E(G_{s,t})\|_2^{k_0}\right\} \\
&\quad +E\left\{\|H_{s,t}-E(H_{s,t})\|_2^{k_0}\right\} \qquad \text{(B.25)}
\end{aligned}
$$

Now from Equation (B.25), we can look at the $E\left\{\|G_{s,t}-E(G_{s,t})\|_2^{k_0}\right\}$ and $E\left\{\|H_{s,t}-E(H_{s,t})\|_2^{k_0}\right\}$. We define the following matrics $L_1=\nabla_{\theta_s}\nabla_{\theta_t}I(\theta_0)$ and $L_2=I^{-1}(\theta_0)$.

$$
\begin{aligned}
E\left\{\|G_{s,t}-E(G_{s,t})\|_2^{k_0}\right\} &\leq 3E[\|\underbrace{\left\{L_1-E(L_1)\right\}L_2}_{F_1}\|_2^{k_0}]+3E[\|\underbrace{E(L_1)\left\{L_2-E(L_2)\right\}}_{F_2}\|_2^{k_0}] \\
&\quad +3\|E\left\{\underbrace{L_1L_2-E(L_1)E(L_2)}_{F_3}\right\}\|_2^{k_0}
\end{aligned}
$$

We look at the bound of F_1. By Jensen's inequality and Holder's inequality, we have

$$
\begin{aligned}
E(\|F_1\|_2^2) &\leq [E\left\{\|L_2\|_2\right\}]^{k_0/2}[E\left\{\|L_1-E(L_1)\|_2\right\}]^{k_0/2} \\
&\leq \mu_-^{k_0}M^{k_0/2}K^{2k_0}\left\{(p+1)\vee\log\alpha_n\right\}^{k_0/2}\alpha_n^{-k_0/2} \qquad \text{(B.26)}
\end{aligned}
$$

Similarly, for F_2, we have

$$\begin{aligned} E(\|F_2\|_2^2) &\leq [E\{\|E(D_1)\|_2\}]^{k_0/2}[E\{\|D_2 - E(D_2)\|_2\}]^{k_0/2} \\ &\leq \|\Upsilon^{-1}\|_2^{k_0/2}\mu_-^{2k_0}M^{k_0/2}K^{k_0}\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2}. \end{aligned} \tag{B.27}$$

For the bound of the bias for F_3, we define

$$U(X_{i1}) = g_2(\beta_0; Z_{i1})X_{i1}X_{i1}^{\mathrm{T}}X_{i1s}X_{i1t}I^{-1}(\theta_0)g_2(\beta_0; Z_{i1})X_{i1}X_{i1}^{\mathrm{T}}I^{-1}(\theta_0) - c_0,$$

where $c_0 = E(D_1)E(D_2)$ and we have

$$\begin{aligned} \|E(F_3)\|_2^{k_0}) &\leq \alpha_n^{-k_0/2}\|E\{U(X_{i1})\}\|_2^{k_0} \\ &\leq \alpha_n^{-k_0/2}E\|U(X_{i1})\|_2^{k_0} \\ &\leq \mu_-^{2k_0}M^{k_0}K^{2k_0}\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2}. \end{aligned} \tag{B.28}$$

Combining the Equation (B.26), Equation (B.27), and Equation (B.28), we have

$$E\left\{\|G_{s,t} - E(G_{s,t})\|_2^{k_0}\right\} \leq C_1^*\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2},$$

where

$$\begin{aligned} C_1^* &= 3\mu_-^{k_0}M^{k_0/2}K^{2k_0} + 3\|\Upsilon^{-1}\|_2^{k_0/2}\mu_-^{2k_0}M^{k_0/2}K^{k_0} \\ &\quad +3\mu_-^{2k_0}M^{k_0}K^{2k_0}. \end{aligned}$$

Similarly, we can define $N_1 = \nabla_{\theta_s}I(\theta_0)$, $N_2 = I^{-1}(\theta_0)$, and $N_3 = \nabla_{\theta_r}I(\theta_0)$.

This would give us

$$
\begin{aligned}
E\left\{\|H_{s,t}-E(H_{s,t})\|_2^{k_0}\right\} \;\leq\; & 5E[\|\underbrace{\{N_1-E(N_1)\}N_2N_3N_2}_{F_1}\|_2^{k_0}] \\
& +5E[\|\underbrace{E(N_1)\{N_2-E(N_2)\}N_3N_2]}_{F_2}\|_2^{k_0}] \\
& +5E[\|\underbrace{E(N_1)E(N_2)\{N_3-E(N_3)\}N_2}_{F_3}\|_2^{k_0}] \\
& +5E[\|\underbrace{E(N_1)E(N_2)E(N_3)\{N_2-E(N_2)\}}_{F_4}\|_2^{k_0}] \\
& +5\|E\left\{\underbrace{N_1N_2N_3N_2-E(N_1)E(N_2)E(N_3)E(N_4)}_{F_5}\right\}\|_2^{k_0}. \qquad \text{(B.29)}
\end{aligned}
$$

We look at the bound of F_1. By Jensen's inequality and Holder's inequality, we have

$$
\begin{aligned}
E(\|F_1\|_2^2) \;&\leq\; [E\{\|N_2\|_2\}]^{k_0}[E\{\|N_3\|_2\}]^{k_0/2}[E\{\|N_1-E(N_1)\|_2\}]^{k_0/2} \\
&\leq\; \mu_-^{3k_0}M^{k_0/2}K^{3k_0/2}\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2} \qquad \text{(B.30)}
\end{aligned}
$$

Similarly, for F_2, we have

$$
\begin{aligned}
E(\|F_2\|_2^2) \;&\leq\; [E\{\|E(N_1)\|_2\}]^{k_0/2}[E\{\|N_2-E(N_2)\|_2\}]^{k_0/2}[E\{\|N_2\|_2\}]^{k_0/2}[E\{\|N_3\|_2\}]^{k_0/2} \\
&\leq\; \|\Theta^{-1}\|_2^{k_0}\mu_-^{3k_0}M^{k_0/2}K^{k_0}\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2}. \qquad \text{(B.31)}
\end{aligned}
$$

For F_3, we have

$$
\begin{aligned}
E(\|F_3\|_2^2) \;&\leq\; [E\{\|E(N_1)\|_2\}]^{k_0/2}[E\{\|N_3-E(N_3)\|_2\}]^{k_0/2}[E\{\|N_2\|_2\}]^{k_0/2}[E\{\|N_2\|_2\}]^{k_0/2} \\
&\leq\; \|\Theta^{-1}\|_2^{k_0/2}\mu_-^{2k_0}M^{k_0/2}K^{3k_0/2}\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2}. \qquad \text{(B.32)}
\end{aligned}
$$

For F_4, we have

$$\begin{aligned} E(\|F_4\|_2^2) &\leq [E\{\|E(N_1)\|_2\}]^{k_0/2}[E\{\|E(N_2)\|_2\}]^{k_0/2}[E\{\|E(N_3)\|_2\}]^{k_0/2} \\ &\quad [E\{\|N_2 - E(N_2)\|_2\}]^{k_0/2} \\ &\leq \|\Theta^{-2}\|_2^{k_0/2}\mu_-^{3k_0}M^{k_0/2}K^{k_0}\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2}. \end{aligned} \tag{B.33}$$

For the bound of the bias for F_5, we define

$$\begin{aligned} U(X_{i1}) &= g_2(\beta_0; Z_{i1})X_{i1}X_{i1}^{\mathrm{T}}X_{i1s}I^{-1}(\theta_0)g_2(\beta_0; Z_{i1})X_{i1}X_{i1}^{\mathrm{T}}I^{-1}(\theta_0) \\ &\quad g_2(\beta_0; Z_{i1})X_{i1}X_{i1}^{\mathrm{T}}X_{i1t}I^{-1}(\theta_0)g_2(\beta_0; Z_{i1})X_{i1}X_{i1}^{\mathrm{T}}I^{-1}(\theta_0) - c_0, \end{aligned}$$

where $c_0 = E(N_1)E(N_2)E(N_3)E(N_4)$ and we have

$$\begin{aligned} \|E(F_5)\|_2^{k_0}) &\leq \alpha_n^{-3k_0/2}\|E\{U(X_{i1})\}\|_2^{k_0} \\ &\leq \alpha_n^{-3k_0/2}E\|U(X_{i1})\|_2^{k_0} \\ &\leq \mu_-^{8k_0}M^{k_0}K^{3k_0}\{(p+1)\vee\log\alpha_n\}^{3k_0/2}\alpha_n^{-3k_0/2}. \end{aligned} \tag{B.34}$$

We use the Equation (B.29) to combine the Equation (B.30), Equation (B.31), Equation (B.32), Equation (B.33), and Equation (B.34). We get

$$E\left\{\|H_{s,t} - E(H_{s,t})\|_2^{k_0}\right\} \leq C_2^*\{(p+1)\vee\log\alpha_n\}^{k_0/2}\alpha_n^{-k_0/2},$$

where

$$\begin{aligned} C_2^* &= 5\mu_-^{3k_0}M^{k_0/2}K^{3k_0/2} + 5\|\Theta^{-1}\|_2^{k_0}\mu_-^{3k_0}M^{k_0/2}K^{k_0} \\ &\quad +5\|\Theta^{-1}\|_2^{k_0/2}\mu_-^{2k_0}M^{k_0/2}K^{3k_0/2} + 5\|\Theta^{-2}\|_2^{k_0/2}\mu_-^{3k_0}M^{k_0/2}K^{k_0} \\ &\quad +5\mu_-^{-8}M^{k_0}K^{3k_0}. \end{aligned}$$

Thus, we have

$$\begin{aligned} E[\|\nabla^2 L_{F,1}(\theta_0) - E\left\{\nabla^2 L_{F,1}(\theta_0)\right\}\|_2^{k_0}] \quad \leq \quad & C_2^{k_0} M^{k_0} K^{k_0} \left\{(p+1) \vee \log \alpha_n\right\}^{k_0/2} \alpha_n^{-k_0/2} \\ & + (p+1)^{k_0+3/2} (2^{k_0}) (C_1^* + C_2^*) \\ & \left\{(p+1) \vee \log \alpha_n\right\}^{3k_0/2} \alpha_n^{-3k_0/2}. \end{aligned}$$

□

B.4. Proof of Lemma 1

We start the proof with the bound of $E(\|\hat{\theta} - \theta_0\|_2^2)$. First, we define the following good events, for the jth site,

$$\begin{aligned} \varepsilon_j = & \left\{\|\hat{\theta} - \theta_0\|_2 \leq \mu_-^{-1} (\phi M)^{1/2} K \Lambda_m\right\} \cap \\ & \left\{\|\nabla L_j(\theta_0)\|_2 \leq (\phi M)^{1/2} K \Lambda\right\} \cap \\ & \left\{\|\nabla^2 L_j(\theta_0) - E\left\{\nabla^2 L_j(\theta_0)\right\}\|_2 \leq M K^2 \Lambda\right\} \cap \\ & \left\{\|\nabla^3 L_j(\theta_0) - E\left\{\nabla^3 L_j(\theta_0)\right\}\|_2 \leq M K^3 \Lambda\right\} \cap \\ & \left\{\|R_4\|_2 \leq M K^4\right\}, \end{aligned} \tag{B.35}$$

where

$$R_{4j} = \int_0^1 \nabla^4 L_j \left\{\theta_0 + v(\hat{\theta}_1 - \theta)\right\} dv$$

,

$$\Lambda = \left[\left\{(p+1) \vee \log \alpha_n\right\} / \alpha_n\right]^{1/2}$$

, and

$$\Lambda_m = \left[\left\{(p+1) \vee \log \alpha_n\right\} / (m \alpha_n)\right]^{1/2}$$

. Then, the good event across sites can be defined as the combination of events, $\varepsilon = \cup_{j=1}^m \varepsilon_j$.

These events describe the closeness between the true parameter θ_0 and the pooled estimator $\hat{\theta}$. We

also defined the following expectation as $\Sigma = E\left\{\nabla^2 L_j(\theta_0)\right\}$, $\Theta = E\left\{\nabla^3 L_j(\theta_0)\right\}$, and $\Upsilon = E\left\{R_{4j}\right\}$. We also defined the expected information matrix as $I = -E\left\{\nabla^2 L_j(\theta_0)\right\}$. We also notice that $E(\nabla^e L_j(\theta_0)$ and $E(\nabla^e L(\theta_0)$ are equal $(e = 1, 2, 3, 4)$.

If the defined events hold, we can use the higher-order Taylor's expansion on the $\nabla L(\theta)$ around the true value θ_0. Denote $\Delta = \hat{\theta} - \theta_0$, we will have,

$$\nabla L(\hat{\theta}_1) = \nabla L(\theta_0) + \nabla^2 L(\theta_0)\Delta + \nabla^3 L(\theta_0)\Delta \otimes \Delta/2 + R_4 \Delta \otimes \Delta \otimes \Delta/6,$$

where $R_4 = \int_0^1 \nabla^4 L\left\{\theta_0 + v(\hat{\theta} - \theta_0)\right\} dv$. With the $\hat{\theta}_1$ as the solution to $\nabla L(\theta) = 0$, we have $\nabla L(\hat{\theta}_1) = 0$. This gives us

$$0 = \nabla L(\theta_0) + \nabla^2 L(\theta_0)\Delta + \nabla^3 L(\theta_0)\Delta \otimes \Delta/2 + R_4 \Delta \otimes \Delta \otimes \Delta/6. \tag{B.36}$$

To further show the higher-order property of the pooled estimator $\hat{\theta}$, we further define the Θ^{-1} by its index notation, shown as $\Theta^{-1}_{i,j,k} = I^{-1}_{i,r} I^{-1}_{j,s} I^{-1}_{k,t} \Theta_{r,s,t}$, which can lower and raise indices. Similarly, we can define Υ^{-1} from its index notation $\Upsilon^{-1}_{i,j,k,l} = I^{-1}_{i,r} I^{-1}_{j,s} I^{-1}_{k,t} I^{-1}_{l,u} \Upsilon_{r,s,t,u}$. Now we can solve the Equation (B.36) by the Equation (7.10) in McCullagh (2018), which gives us

$$\begin{aligned} \Delta &= I^{-1}\nabla L(\theta_0) + I^{-2}\left\{\nabla^2 L(\theta_0) - \Sigma\right\}\nabla L(\theta_0) + \Theta^{-1}\nabla L(\theta_0) \otimes \nabla L(\theta_0)/2 \\ &\quad + O(1) I^{-1}\nabla L(\theta_0) \otimes \nabla L(\theta_0) \otimes \nabla L(\theta_0)/6. \end{aligned} \tag{B.37}$$

Here to preserve the simplicity of the Equation (B.37), we use the $O(1)$ to indicate the five items omitted, which has the same order as $1/6\Upsilon^{-1}\nabla L(\theta_0) \otimes \nabla L(\theta_0) \otimes \nabla L(\theta_0)$. Thus, these five items along with $1/6\Upsilon^{-1}\nabla L(\theta_0) \otimes \nabla L(\theta_0) \otimes \nabla L(\theta_0)$ would have the same bound. The details of showing the bound of the mean square error and bias of these items will be shown in the Lemma B.3.12.

The rest of the proof is to find the bounds of $\|E(\Delta)\|_2^2$ and $E(\|\Delta\|_2^2)$.

Now for the first part of the proof, define a random variable $I(\varepsilon)$ as the indicator function whether

the event ε happens. We denote $B_1 = I^{-1}\nabla L(\theta_0)$, $B_2 = I^{-2}(\nabla^2 L(\theta_0) - \Sigma)\nabla L(\theta_0)$, $B_3 = 1/2\Theta^{-1}\nabla L(\theta_0) \otimes \nabla L_1(\theta_0)$, and $B_4 = O(1)1/6\Upsilon^{-1}\nabla L(\theta_0) \otimes \nabla L(\theta_0) \otimes \nabla L(\theta_0)$.

For the item $\|E(\Delta)\|_2^2$, we have

$$\begin{aligned}
\|E(\Delta)\|_2^2 &\leq \|E\left\{I(\varepsilon)\Delta\right\}\|_2^2 + P(\varepsilon^c) \\
&\leq \|E\left\{B_1 + B_2 + B_3 + B_4\right\}\|_2^2 + P(\varepsilon^c) \\
&\leq 4\|E(B_1)\|_2^2 + 4\|E(B_2)\|_2^2 + 4\|E(B_3)\|_2^2 + 4\|E(B_4)\|_2^2 \\
&\quad +12m\alpha_n^{-8} + 2mn^{-8}
\end{aligned} \tag{B.38}$$

The probability of event ε that does not hold can be found by combining the event probability for each site and using the Lemma B.3.9, Lemma B.3.1, Lemma B.3.2, Lemma B.3.3, and Lemma B.3.4.

Now we investigate the bound of the bias from B_1 to B_4 when the event holds. First, we notice that $E\left\{\nabla L_1(\theta_0)\right\} = 0$ and thus, we have

$$\|E(B_1)\|_2^2 = 0. \tag{B.39}$$

For B_2, we first define the functions $U(X) = g_2(Z;\beta_0)XX^T - \Sigma$, and then we can rewrite $E(B_2)$ as

$$\begin{aligned}
E(B_2) &= E\left\{I^{-2}\left\{\nabla^2 L(\theta_0) - \Sigma\right\}\nabla L(\theta_0)\right\} \\
&= (m\alpha_n)^{-2}E[\sum_{j=1}^{m}\sum_{i=1}^{n} I^{-2}\left\{\nabla^2 L(\theta_0; X_{ij}) - \Sigma\right\}\nabla L(\theta_0; X_{ij})] \\
&= (m\alpha_n)^{-1}E[I^{-2}U(X_{ij})\nabla L(\theta_0; X_{ij})]
\end{aligned}$$

The last equality holds from the Lemma B.3.2.

Then, by the Jensen's inequality and Cauchy-Schwarz inequality we have

$$\begin{aligned}
\|E(B_2)\|_2^2 &= (m\alpha_n)^{-2}\|E\left\{I^{-2}U(X_{i1})\nabla L(\theta_0;X_{ij})\right\}\|_2^2 \\
&\leq (m\alpha_n)^{-2}(E[\|I^{-2}U(X_{i1})\nabla L(\theta_0;X_{ij})\|_2])^2 \\
&\leq (m\alpha_n)^{-2}\mu_-^{-4}E\left\{\|U(X_{i1})\|_2^2\right\}E\left\{\|\nabla L(\theta;X_{ij})\|_2^2\right\}
\end{aligned}$$

Notice that

$$E\left\{\|U(X_{i1})\|_2^2\right\} \leq M^2K^4(p+1)^2$$

and

$$E\left\{\|\nabla L_1(\theta;X_{i1})\|_2^2\right\} \leq \left\{E\|X_{i1}\|_2^4 E\|Y_{i1} - \frac{g_1(\alpha_0)g_2(X_{i1};\beta_0)}{\{1+g_1(\alpha_0)g_2(X_{i1};\beta_0)\}^2}\|_2^4\right\}^{1/2} \leq \phi M K^4.$$

Then, we have

$$\|E(B_2)\|_2^2 = \mu_-^{-4}\phi M^3K^8\left\{\frac{p+1}{m\alpha_n}\right\}^2. \tag{B.40}$$

For the expectation of B_3, we can use the Jensen's inequality and Cauchy-Schwarz inequality to get

$$\begin{aligned}
\|E(B_3)\|_2^2 &= (1/4)\|E\left\{\Theta^{-1}\nabla L(\theta_0)\otimes\nabla L(\theta_0)\right\}\|_2^2 \\
&= (1/4)\|(m\alpha_n)^{-1}\Theta^{-1}E\left\{-\nabla^2 L(\theta_0;X_{ij})\right\}\|_2^2 \\
&\leq (1/4)(m\alpha_n)^{-2}[E\left\{\|\Theta^{-1}\nabla^2 L(\theta_0;X_{ij})\|_2\right\}]^2 \\
&\leq (1/4)(m\alpha_n)^{-2}\|\Theta^{-1}\|_2^2 E[\|\left\{\nabla^2 L(\theta_0;X_{ij})\right\}\|_2^2] \\
&\leq (1/4)(m\alpha_n)^{-2}\|\Theta^{-1}\|_2^2 M^2K^4(p+1)^2.
\end{aligned} \tag{B.41}$$

The second equality is from the second Bartlett identity $E\left\{\sum_{j=1}^m\sum_{i=1}^n \nabla L(\theta_0;X_{ij})\otimes L(\theta_0;X_{ij})\right\} = -E\left\{\sum_{j=1}^m\sum_{i=1}^n \nabla^2 L(\theta_0;X_{i1})\right\}$.

As for the expectation of B_4, we have

$$\|E(B_4)\|_2^2 \quad = \quad C''(p+1)^4(m\alpha_n)^{-4}, \tag{B.42}$$

where

$$\begin{aligned} C'' \quad &= \quad 6\mu_-^{-6}\phi M^5K^{10} + 6\|\Theta^{-1}\|_2^2\mu_-^{-2}\phi^2M^4K^8 + (3/2)\mu_-^{-4}\phi^2M^4K^8 \\ &\quad +(3/2)\mu_+^2\mu_-^{-4}M^2K^6 + (3/2)\mu_-^{-6}M^4K^{10} + (1/6)\|\Upsilon^{-1}\|_2^2M^2K^6. \end{aligned}$$

The details of bias of full items for B_4 is shown by applying the similar arguments from the Lemma B.3.12 to the pooled data.

Combining the results from Equation (B.39), Equation (B.40), Equation (B.41), and Equation (B.42), we have

$$\begin{aligned} \|E\{\Delta\}\|_2^2 \quad &\leq \quad (C_1'' + C_2'')\frac{\{(p+1) \vee 8\log\alpha_n)\}^2}{(m\alpha_n)^2} + C''\frac{\{(p+1) \vee 8\log\alpha_n)\}^4}{(m\alpha_n)^4} \\ &\quad +12m\alpha_n^{-8} + 2mn^{-8}, \end{aligned} \tag{B.43}$$

where $C_1'' = \mu_-^{-4}\phi M^3K^8$, $C_2'' = \|\Theta^{-1}\|_2^2M^2K^4$.

From Equation (B.43), we find the dominated term is related to $(C_1''+C_2'')\{(p+1) \vee 8\log\alpha_n)\}^2(m\alpha_n)^{-2}$, so we can denote $C_1 = C_1'' + C_2''$ and we have $\|E(\Delta)\|_2^2 \leq C_1\{(p+1) \vee 8\log\alpha_n)\}^2(m\alpha_n)^{-2}$. This completes the first part of the proof.

Then, to find the bound of $E(\|\Delta\|_2^2)$. Notice that

$$\begin{aligned} E(\|\Delta\|_2^2) \quad &= \quad E\left\{I(\varepsilon)\|\Delta\|_2^2\right\} + E\left\{I(\varepsilon^c)\|\Delta\|_2^2\right\} \\ &\leq \quad E\left\{I(\varepsilon)\|\Delta\|_2^2\right\} + pr(\epsilon^c) \\ &\leq \quad E\left\{I(\varepsilon)\|\Delta\|_2^2\right\} + \sum_{t=0}^{4} pr(\epsilon_t^c) \\ &= \quad E\left\{I(\varepsilon)\|\Delta\|_2^2\right\} + 12m\alpha_n^{-8} + 2mn^{-8}. \end{aligned} \tag{B.44}$$

Now under the event ε, we investigate Δ.

$$\begin{aligned} E\left\{I(\varepsilon)\|\Delta\|_2^2\right\} &\leq E\left\{I(\varepsilon)\|B_1+B_2+B_3+B_4\|_2^2\right\} \\ &\leq 4E(I(\varepsilon)\|B_1\|_2^2)+4E(I(\varepsilon)\|B_2\|_2^2)+4E(I(\varepsilon)\|B_3\|_2^2) \\ &\quad +4E(I(\varepsilon)\|B_4\|_2^2). \end{aligned} \tag{B.45}$$

By using the Lemma B.3.1 and Cauchy-Schwarz inequality, we have the bound for B_1,

$$\begin{aligned} E\left\{I(\varepsilon)\|B_1\|_2^2\right\} &\leq \|I^{-1}\|_2^2 E(\|\nabla L(\theta_0)\|_2^2) \\ &\leq \mu_-^{-2}(\phi M)^{1/2}K\left\{(p+1)\vee\log\alpha_n\right\}(m\alpha_n)^{-1}. \end{aligned} \tag{B.46}$$

Also we have the bound for B_2,

$$\begin{aligned} E(I(\varepsilon)\|B_2\|_2^2) &\leq \|I^{-2}\|_2^2 E\left\{\|(\nabla^2 L(\theta_0)-\Sigma)\nabla L(\theta_0)\|_2^2\right\} \\ &\leq \|I^{-2}\|_2^2[E\left\{\|(\nabla^2 L(\theta_0)-\Sigma)\|_2^4\right\}E\left\{\|\nabla L(\theta_0)\|_2^4\right\}]^{1/2} \\ &\leq \mu_-^{-2}\phi M^3K^6\left\{(p+1)\vee\log\alpha_n\right\}^2(m\alpha_n)^{-2}. \end{aligned} \tag{B.47}$$

The first inequality and the second inequality are by using the Cauchy-Schwarz inequality. The last one is by using the Lemma B.3.1 and the Lemma B.3.2.

For B_3, by using the Lemma (B.3.1), we have

$$\begin{aligned} E(I(\varepsilon)\|B_3\|_2^2) &\leq (1/4)\|\Theta^{-1}\|_2^2 E\left\{\|\nabla L(\theta_0)\otimes\nabla L(\theta_0)\|_2^2\right\} \\ &\leq (1/4)\|\Theta^{-1}\|_2^2 E\left\{\|\nabla L(\theta_0)\|_2^4\right\} \\ &\leq (1/4)\|\Theta^{-1}\|_2^2\phi^2M^2K^2\left\{(p+1)\vee\log\alpha_n\right\}^2(m\alpha_n)^{-2}. \end{aligned} \tag{B.48}$$

As applying the Lemma B.3.11 to the pooled data, we have

$$E(I(\varepsilon)\|B_4\|_2^2) \leq C'\left\{(p+1)\vee\log\alpha_n\right\}^3(m\alpha_n)^{-3}, \tag{B.49}$$

where

$$\begin{aligned} C' &= 6\mu_-^{-6}\phi M^5K^{10} + 6\mu_-^{-2}\|\Theta^{-1}\|_2^2\phi^2M^4K^8 + (3/2)\mu_-^{-2}\|\Theta^{-1}\|_2^2\phi^2M^4K^8 + (3/2)\mu_+^2\|\Theta^{-2}\|_2^2\phi^3M^3K^6 \\ &\quad +(3/2)\mu_-^{-6}\phi^2M^4K^{10} + (1/6)\|\Upsilon^{-1}\|_2^2\phi^3M^3K^6. \end{aligned}$$

Combining the results from Equation (B.46), Equation (B.47), Equation (B.48), and Equation (B.49) by using the Equation (B.45), we have

$$\begin{aligned} E\left\{I(\varepsilon)\|\Delta\|_2^2\right\} &\leq C_1'\frac{(p+1)\vee 8\log\alpha_n}{m\alpha_n} + (C_2'+C_3')\left\{\frac{(p+1)\vee 8\log\alpha_n}{m\alpha_n}\right\}^2 \\ &\quad +C'\left\{\frac{(p+1)\vee 8\log\alpha_n}{m\alpha_n}\right\}^3, \end{aligned} \tag{B.50}$$

where $C_1' = \mu_-^{-2}(\phi M)^{1/2}K$, $C_2' = \mu_-^{-2}\phi M^3K^6$, and $C_3' = (1/4)\|\Theta^{-1}\|_2^2\phi^2M^2K^2$.

From Equation (B.50), we find the dominated term is related to $C_1'\frac{(p+1)\vee 8\log\alpha_n}{m\alpha_n}$, so we can denote $C_2 = C_1' + C_2' + C_3' + C'$ and $E\left\{I(\varepsilon)\|\Delta\|_2^2\right\} \leq C_2(m\alpha_n)^{-1}(p+1)\vee 8\log\alpha_n$.

From Equation (B.44), we have

$$E(\|\Delta\|_2^2) \leq C_2\frac{p\vee 4\log\alpha_n}{\alpha_n} + 12m\alpha_n^{-8} + 2mn^{-8}. \tag{B.51}$$

This completes the second part of the proof related to the bound of the mean square error for the pooled estimator.

B.5. Proof of Lemma 2

To identify the systematic bias of the distributed algorithms, we used the Equation (24) in Zhang et al. (2012), which gives us

$$E(\|\bar{\theta} - \theta_0\|_2^2) \leq m^{-1} E(\|\hat{\theta}_1 - \theta_0\|_2^2) + \|E(\hat{\theta}_1 - \theta_0)\|_2^2. \tag{B.52}$$

From the Equation (B.52), the error bound of the mean square error of $\bar{\theta}$ can be seen as the sum of two parts, which are the mean square error of a local site divided by the number of sites and the expected bias for a local site. The second part is also called the systematic bias of the distributed estimators as it is related to the performance of a local site. Our major contribution is to find the error by studying the behavior of a local estimator under the rare disease setting. The assumptions listed above gives conditions for the smoothness of the first to the fourth order derivatives.

We first define the following good events

$$\begin{aligned}
\varepsilon_0 &= \left\{ \|\hat{\theta}_1 - \theta_0\|_2 \leq \mu_-^{-1} (\phi M)^{1/2} K\Lambda \right\}, \\
\varepsilon_1 &= \left\{ \|\nabla L_1(\theta_0)\|_2 \leq (\phi M)^{1/2} K\Lambda \right\}, \\
\varepsilon_2 &= \left\{ \|\nabla^2 L_1(\theta_0) - E\left\{\nabla^2 L_1(\theta_0)\right\}\|_2 \leq MK^2\Lambda \right\}, \\
\varepsilon_3 &= \left\{ \|\nabla^3 L_1(\theta_0) - E\left\{\nabla^3 L_1(\theta_0)\right\}\|_2 \leq MK^3\Lambda \right\}, \\
\varepsilon_4 &= \left\{ \|R_4\|_2 \leq MK^4 \right\},
\end{aligned} \tag{B.53}$$

where $R_4 = \int_0^1 \nabla^4 L_1 \left\{\theta_0 + v(\hat{\theta}_1 - \theta)\right\} dv$ and $\Lambda = [\{(p+1) \vee \log \alpha_n\} / \alpha_n]^{1/2}$.

These events describe the closeness between the true parameter θ_0 and local estimator $\hat{\theta}_1$. We define a new event ε as the intersection of event ε_1 to ε_4, that is, $\varepsilon = \bigcap_{t=0}^4 \varepsilon_t$. We also defined the following expectation as $\Sigma = E\left\{\nabla^2 L_1(\theta_0)\right\}$, $\Theta = E\left\{\nabla^3 L_1(\theta_0)\right\}$, and $\Upsilon = E\left\{R_4\right\}$. We also defined the expected information matrix as $I = -E\left\{\nabla^2 L_1(\theta_0)\right\}$ If the defined events hold, we can use the higher-order Taylor's expansion on the $\nabla L_1(\theta)$ around the true value θ_0. Denote $\Delta = \hat{\theta}_1 - \theta_0$, we

will have,

$$\nabla L_1(\hat{\theta}_1) = \nabla L_1(\theta_0) + \nabla^2 L_1(\theta_0)\Delta + \nabla^3 L_1(\theta_0)\Delta \otimes \Delta/2 + R_4 \Delta \otimes \Delta \otimes \Delta/6.$$

With the $\hat{\theta}_1$ as the solution to $\nabla L_1(\theta) = 0$, we have $\nabla L_1(\hat{\theta}_1) = 0$. This gives us

$$0 = \nabla L_1(\theta_0) + \nabla^2 L_1(\theta_0)\Delta + \nabla^3 L_1(\theta_0)\Delta \otimes \Delta/2 + R_4 \Delta \otimes \Delta \otimes \Delta/6. \tag{B.54}$$

The rest of the proof can be followed from the similar arguments for the proof of the Lemma 1 in the main manuscript by applying the Lemmas on the local site.

Then we have

$$\begin{aligned} E(\|\bar{\theta} - \theta_0\|_2^2) \quad \leq \quad & m^{-1} C_3 \alpha_n^{-1} \left\{(p+1) \vee 8 \log \alpha_n\right\} + C_4 \left\{(p+1) \vee 8 \log \alpha_n\right\}^2 \alpha_n^{-2} \\ & + 12 \alpha_n^{-8} + 2n^{-8}, \end{aligned} \tag{B.55}$$

where $C_3 = C_2$ and $C_4 = C_1$.

This completes the proof.

B.6. Proof of Theorem 1

To begin the proof of the Theorem 1, we first show the bound of the one-round distributed Firth estimator $\tilde{\theta}^{(1)}$. Later we show the bound of the two-round distributed Firth estimator $\tilde{\theta}^{(2)}$.

For $\tilde{\theta}^{(1)}$, we follow the definition of events shown in the Equation (B.53). This is because the one-round distributed Firth estimator can be seen as the average of the local Firth-corrected estimator across different sites. In other words, the following equation holds,

$$E(\|\tilde{\theta}^{(1)} - \theta_0\|_2^2) \leq m^{-1} E(\|\hat{\theta}_{F,1} - \theta_0\|_2^2) + \|E(\hat{\theta}_{F,1} - \theta_0)\|_2^2. \tag{B.56}$$

Thus, we can first prove the bound of $\|E(\hat{\theta}_{F,1} - \theta_0)\|_2^2$ and $E(\|\hat{\theta}_{F,1} - \theta_0\|_2^2)$. Then, we can use

Equation (B.56) to finish the proof.

We define the following good event with respect to $\hat{\theta}_{F,1}$,

$$\varepsilon_{F,0} = \left\{ \|\hat{\theta}_{F,1} - \theta_0\|_2 \leq C_1 \mu_-^{-1} (\phi M)^{1/2} K\Lambda \right\}.$$

And we define the event $\varepsilon_F = \varepsilon_{F,0} \cap \varepsilon_1 \cap \varepsilon_2 \cap \varepsilon_3 \cap \varepsilon_4$.

When the event ε_F holds, we can use the higher-order Taylor's expansion on the $\nabla L_{F,1}(\theta)$ around the true value θ_0. Denote $\Delta_F = \hat{\theta}_{F,1} - \theta_0$, we will have,

$$\nabla L_{F,1}(\hat{\theta}_{F,1}) = \nabla L_1(\theta_0) + \nabla^2 L_1(\theta_0)\Delta + \nabla^3 L_1(\theta_0)\Delta \otimes \Delta/2 + R_4 \Delta \otimes \Delta \otimes \Delta/6 + T(\theta_0),$$

where the rth element of T is $T_r(\theta_0) = \alpha_n^{-1} b_r(\theta_0)$.

With the $\hat{\theta}_{F,1}$ as the solution to $\nabla L_{F,1}(\theta) = 0$, we have $\nabla L_{F,1}(\hat{\theta}_1) = 0$. This gives us

$$0 = \nabla L_1(\theta_0) + \nabla^2 L_1(\theta_0)\Delta + \nabla^3 L_1(\theta_0)\Delta \otimes \Delta/2 + R_4 \Delta \otimes \Delta \otimes \Delta/6 + T(\theta_0). \tag{B.57}$$

By solving the Equation (B.57), we have

$$\begin{aligned} \Delta_F &= I^{-1}\nabla L_1(\theta_0) + I^{-2}\left\{\nabla^2 L_1(\theta_0) - \Sigma\right\}\nabla L_1(\theta_0) + \Theta^{-1}\nabla L_1(\theta_0) \otimes \nabla L_1(\theta_0)/2 \\ &\quad + O(1)\Upsilon^{-1}\nabla L_1(\theta_0) \otimes \nabla L_1(\theta_0) \otimes \nabla L_1(\theta_0)/6 + I^{-1}T(\theta_0). \end{aligned} \tag{B.58}$$

The rest of the proof for the one-round distributed Firth estimator would be to find the bound of $E(\|\Delta_F\|_2^2)$ and $\|E(\Delta_F)\|_2^2$ with the help of Equation (B.58).

Notice that

$$E(\|\Delta_F\|_2^2) = E\left\{I(\varepsilon)\|\Delta_F\|_2^2\right\} + 12\alpha_n^{-8} + 2n^{-8}. \tag{B.59}$$

We define $B_5 = I^{-1}\alpha_n^{-1}b(\theta_0)$, B_1 to B_4 are defined in the Section of proof of the lemma 2.

We shall have

$$\begin{aligned} E\left\{I(\varepsilon_F)\|\Delta\|_2^2\right\} &\leq E\left\{I(\varepsilon)\|B_1+B_2+B_3+B_4+B_5\|_2^2\right\} \\ &\leq 5E(I(\varepsilon)\|B_1\|_2^2)+5E(I(\varepsilon)\|B_2\|_2^2)+5E(I(\varepsilon)\|B_3\|_2^2) \\ &\quad +5E(I(\varepsilon)\|B_4\|_2^2)+5E(I(\varepsilon)\|B_5\|_2^2). \end{aligned} \tag{B.60}$$

For B_5, we can have the following results,

$$\begin{aligned} E\left\{I(\varepsilon_F)\|B_5\|_2^2\right\} &\leq (1/4)\|I^{-1}\|_2^2 E(\|\alpha_n^{-1}b(\theta_0)\|_2^2) \\ &\leq (1/4)\|I^{-1}\|_2^2\alpha_n^{-2}\left\{E(\|I^{-1}(\theta_0)\|_2^4)E(\|\nabla I(\theta_0)\|_2^4)\right\}^{1/2} \\ &\leq (1/4)\mu_-^{-6}M^4K^{10}(p\vee\log\alpha_n)^2\alpha_n^{-2}. \end{aligned} \tag{B.61}$$

The last inequality is using the Lemma B.3.3 and Cauchy-Schwarz inequality, we have the bound for B_5.

From Equation (B.46), Equation (B.47), Equation (B.48), and Equation (B.49) and Equation (B.61) by using the Equation (B.60), we have

$$\begin{aligned} E\left\{I\left(\varepsilon\right)\|\Delta\|_2^2\right\} &\leq (C_{F,1})\frac{p\vee 4\log\alpha_n}{\alpha_n}+(C_{F,2}+C_{F,3}+C_{F,4})\left\{\frac{p\vee 4\log\alpha_n}{\alpha_n}\right\}^2 \\ &\quad +C_{F,5}\left\{\frac{p\vee 4\log\alpha_n}{\alpha_n}\right\}^3, \end{aligned} \tag{B.62}$$

where

$$\begin{aligned} C_{F,1} &= 5\mu_-^{-2}(\phi M)^{1/2}K, \\ C_{F,2} &= 5\mu_-^{-4}\phi M^3 K^6, \\ C_{F,3} &= (5/4)\mu_-^{-6}M^4K^{10}, \\ C_{F,4} &= 5/4\|\Theta^{-1}\|_2^2\phi^2M^2, \\ C_{F,5} &= O(1)(5/36)\|\Upsilon^{-1}\|_2^2(\phi M)^3K^6. \end{aligned}$$

Then, we look at the bias term of B_5. As shown in the Lemma B.3.13, we have

$$\|E(B_2+B_3+B_5)\|_2^2 \leq \mu_-^{-6}M^4K^{10}p^4\alpha_n^{-4}.$$

Combing the information that $E(B_1) = E\{\nabla L(\theta_0)\} = 0$, we may get

$$\|E(\Delta_F)\|_2^2 = 2\|E(B_4)\|_2^2 + 2\|E(B_2+B_3+B_5)\|_2^2 + 12\alpha_n^{-8} + 2n^{-8}. \tag{B.63}$$

From Lemma B.3.12, this shows that

$$\|E(\Delta_F)\|_2^2 = C_{F,6}\frac{\{(p+1)\vee 4\log\alpha_n\}}{\alpha_n^4} + 12\alpha_n^{-8} + 2n^{-8}.$$

By using the Equation (B.56) this will give us

$$E(\|\tilde{\theta}^{(1)}-\theta_0\|_2^2) \leq C_5 m^{-1}\{(p+1)\vee 8\log\alpha_n\}\alpha_n^{-1} + C_6\{(p+1)\vee 8\log\alpha_n\}^4\alpha_n^{-4}, \tag{B.64}$$

where $C_5 = C_{F,1}+C_{F,2}+C_{F,3}+C_{F,4}+C_{F,5}$, and $C_6 = C_{F,6}$.

This finishes the proof related to the one-round distributed Firth estimator.

For the two-round distributed Firth estimator, we majorly follow the proof stated in Jordan et al.

(2018). We define the good events for jth site ($j = 1, .., m$),

$$\begin{aligned}\varepsilon_j &= \left\{\|\hat{\theta}_{F,j} - \theta_0\|_2 \leq \mu_-^{-1}(\phi M)^{1/2} K\Lambda\right\} \cap \\ &\quad \left\{\|\nabla L_j(\theta_0)\|_2 \leq (\phi M)^{1/2} K\Lambda\right\} \cap \\ &\quad \left\{\|R_{4,j}\|_2 \leq MK^4\right\} \cap \\ &\quad \left\{\|\nabla^2 L_j(\theta_0) - E\left\{\nabla^2 L_j(\theta_0)\right\}\|_2 \leq MK^2\Lambda\right\} \cap \\ &\quad \left\{\|\nabla^3 L_j(\theta_0) - E\left\{\nabla^3 L_j(\theta_0)\right\}\|_2 \leq MK^3\Lambda\right\},\end{aligned}$$

where $R_{4,j} = \int_0^1 \nabla^4 L_1 \left\{\theta_0 + v(\hat{\theta}_1 - \theta)\right\} dv$.

When the event does not hold, from Lemma B.3.1, Lemma B.3.2, Lemma B.3.3, Lemma B.3.4, and Lemma B.3.6, we have

$$P(\varepsilon^c) = \frac{12m}{\alpha_n^8} + \frac{2m}{n^8}.$$

Under the event $\varepsilon = \cup_{j=1}^m \varepsilon_j$, from Lemma B.3.14 we have

$$\|\tilde{\theta}^{(2)} - \hat{\theta}_F\|_2 \leq 4\mu_-^{-1}\|\nabla \tilde{L}_F(\hat{\theta})\|_2.$$

Now we are looking at the term $\nabla \tilde{L}_F(\hat{\theta})$. Notice that $\nabla L_F(\hat{\theta}_F) = 0$, and thus we have

$$\nabla \tilde{L}(\hat{\theta}_F) = \nabla L_{F,1}(\hat{\theta}_F) - \nabla L_{F,1}(\tilde{\theta}^{(1)}) - \left\{\nabla L_F(\hat{\theta}_F) - \nabla L_F(\tilde{\theta}^{(1)})\right\}. \tag{B.65}$$

We use the Taylor's Expansion on the $\nabla L_{F,1}(\theta)$ around the $\nabla L_{F,1}(\tilde{\theta}^{(1)})$, which gives us

$$\nabla L_{F,1}(\hat{\theta}_F) = \nabla L_{F,1}(\tilde{\theta}^{(1)}) + P_1(\hat{\theta}_F - \tilde{\theta}^{(1)}),$$

where $P_1 = \int_0^1 \nabla^2 L_{F,j} \left\{\tilde{\theta}^{(1)} + v(\hat{\theta}_F - \tilde{\theta}^{(1)})\right\} dv$.

Notice that $\nabla^2 L_{F,1}(\theta) = \nabla^2 L_1(\theta) + \nabla T(\theta)$, where the sth row and tth column of $\nabla T(\theta)$ is the sum of traces for matrices $G_{s,t} = I^{-1}(\theta_0)\nabla_{\theta_s}\nabla_{\theta_t} I(\theta_0)$ and $H_{s,t} = \nabla_{\theta_s} I(\theta_0) I^{-1}(\theta_0)\nabla_{\theta_t} I(\theta_0) I^{-1}(\theta_0)$. As

shown in the Lemma B.3.6, $I(\theta_0)$, $\nabla I(\theta_0)$, and $\nabla^2 I(\theta_0)$ are Lipschitz continuous and thus $\nabla^2 L_{F,1}(\theta)$ is Lipschitz. This gives us

$$\|P_1 - \nabla L_{F,1}(\theta_0)\|_2 \leq M'(\|\hat{\theta}_F - \theta_0\|_2 + \|\tilde{\theta}^{(1)} - \theta_0\|_2).,$$

where M' is a constant. Similarly, we have for $\nabla L_F(\theta)$,

$$\nabla L_F(\hat{\theta}_F) = \nabla L_F(\tilde{\theta}^{(1)}) + P(\hat{\theta}_F - \tilde{\theta}^{(1)}),$$

where $P = \int_0^1 \nabla^2 L_F \left\{\tilde{\theta}^{(1)} + v(\hat{\theta}_F - \tilde{\theta}^{(1)})\right\} dv$ satisfies

$$\|P - \nabla L_F(\theta_0)\|_2 \leq M'(\|\hat{\theta}_F - \theta_0\|_2 + \|\tilde{\theta}^{(1)} - \theta_0\|_2).$$

From Equation (B.65), we have

$$\begin{aligned}
\|\nabla \tilde{L}(\hat{\theta}_F)\|_2 &= (\|P_1 - \nabla L_{F,1}(\theta_0)\|_2 + \|\nabla L_{F,1}(\theta_0) - \nabla L_F(\theta_0)\|_2 + \|P - \nabla L_F(\theta_0)\|_2) \\
&\quad \|\hat{\theta}_F - \tilde{\theta}^{(1)}\|_2. \\
&\leq (2M'\|\hat{\theta}_F - \theta_0\|_2 + 2M'\|\tilde{\theta}^{(1)} - \theta_0\|_2 + \|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2) \\
&\quad \|\hat{\theta}_F - \tilde{\theta}^{(1)}\|_2 \\
&\leq 4M'\|\hat{\theta}_F - \theta_0\|_2^2 + 4M'\|\tilde{\theta}^{(1)} - \theta_0\|_2^2 + \|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2\|\hat{\theta}_F - \theta_0\|_2 \\
&\quad + \|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2\|\tilde{\theta}^{(1)} - \theta_0\|_2 \qquad \text{(B.66)}
\end{aligned}$$

Notice that from the triangular inequality, we have

$$\begin{aligned}
\|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2 &\leq 3\|\nabla L_F(\theta_0) - E\left\{\nabla L_F(\theta_0)\right\}\|_2 + 3\|E\left\{\nabla L_F(\theta_0)\right\} - E\left\{\nabla L_{F,1}(\theta_0)\right\}\|_2 \\
&\quad +3\|\nabla L_{F,1}(\theta_0) - E\left\{\nabla L_{F,1}(\theta_0)\right\}\|_2 \\
&= 3\|\nabla L_F(\theta_0) - E\left\{\nabla L_F(\theta_0)\right\}\|_2 + 3\|E\left\{T(\theta_0)\right\} - E\left\{T_1(\theta_0)\right\}\|_2 \\
&\quad +3\|\nabla L_{F,1}(\theta_0) - E\left\{\nabla L_{F,1}(\theta_0)\right\}\|_2 \\
&\leq 3\|\nabla L_F(\theta_0) - E\left\{\nabla L_F(\theta_0)\right\}\|_2 + 3\|\nabla L_{F,1}(\theta_0) - E\left\{\nabla L_{F,1}(\theta_0)\right\}\|_2 \\
&\quad 6\|E\left\{T(\theta_0)\right\}\|_2 + 6\|E\left\{T_1(\theta_0)\right\}\|_2. \qquad \text{(B.67)}
\end{aligned}$$

With the help of the Equation (B.66), we have

$$\begin{aligned}
E\left\{\|\tilde{\theta}^{(2)} - \hat{\theta}_F\|_2^2 I(\varepsilon)\right\} &\leq 4\mu_-^{-1} E\left\{\|\nabla \tilde{L}(\hat{\theta}_F)\|_2^2\right\} \\
&\leq 64\mu_-^{-1} M'\left\{E(\|\hat{\theta}_F - \theta_0\|_2^4) + E(\|\tilde{\theta}^{(1)} - \theta_0\|_2^4)\right\} \\
&\quad +16\mu_-^{-1} E[\|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2^2 \|\hat{\theta}_F - \theta_0\|_2^2] \\
&\quad +16\mu_-^{-1} E[\|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2^2 \|\tilde{\theta}^{(1)} - \theta_0\|_2^2] \\
&\leq 64\mu_-^{-1} M'\left\{E(\|\hat{\theta}_F - \theta_0\|_2^4) + E(\|\tilde{\theta}^{(1)} - \theta_0\|_2^4)\right\} \\
&\quad +16\mu_-^{-1}[E\left\{\|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2^4\right\}]^{1/2}\left\{E(\|\hat{\theta}_F - \theta_0\|_2^4)\right\}^{1/2} \\
&\quad +16\mu_-^{-1}[E\left\{\|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2^4\right\}]^{1/2}\left\{E(\|\tilde{\theta}^{(1)} - \theta_0\|_2^4)\right\}^{1/2}.
\end{aligned}$$

From Lemma B.3.13 and Lemma B.3.15, we have

$$E\left\{\|\nabla L_F(\theta_0) - \nabla L_{F,1}(\theta_0)\|_2^4\right\} \leq C_1^*[\frac{\{(p+1) \vee \log(\alpha_n)\}^2}{(K\alpha_n)^2} + \frac{\{(p+1) \vee \log(\alpha_n)\}^2}{\alpha_n^2}], \qquad \text{(B.68)}$$

where

$$C_1^* = 3(2^4)(\phi M)^2K^2 + 9(2^4)(p+1)^3\mu_-^{-2}M^{k_0/2}K^6 + 9(2^4)(p+1)^3 \\ \|\Theta^{-1}\|_2^2\mu_-^{-2}M^2K^4 + 9(2^4)(p+1)^3\mu_-^{-4}M^4K^{10} + 6\mu_-^{-12}M^8K^{20}.$$

From the Lemma 2 in the main manuscript, we have

$$E\|(\tilde{\theta}^{(1)} - \theta_0)\|_2^4 \leq \left\{\frac{C_5(p+1) \vee \log \alpha_n}{m\alpha_n}\right\}^2 + \frac{C_6^2\left\{(p+1) \vee \log \alpha_n\right\}^8}{\alpha_n^8},$$

and

$$E\|(\hat{\theta}_F - \theta_0)\|_2^4 \leq \left\{\frac{C_1(p+1) \vee \log \alpha_n}{m\alpha_n}\right\}^2.$$

This would give us,

$$E\left\{\|\tilde{\theta}^{(2)} - \theta_0\|_2^2\right\} \leq E\left\{\|\tilde{\theta}^{(2)} - \hat{\theta}_F\|_2^2\right\} + E\left\{\|\hat{\theta}_F - \theta_0\|_2^2\right\} \leq C_7\frac{(p+1) \vee \log \alpha_n}{m\alpha_n} + C_8\frac{\left\{(p+1) \vee \log \alpha_n\right\}^5}{\alpha_n^5},$$

where $C_7 = 16\mu_-^{-1}\left\{C_1 + C_5 + (C_1^*)^{1/2}\right\}$, and $C_8 = 16\mu_-^{-1}\left\{(C_1^*)^{1/2}C_6\right\}$.

This completes the proof.

B.7. Remark 1

Remark 2. We can extend all the above results to the unequal sample size scenario. Specifically, for Theorem 1, if we denote the sample size at each site as n_j, $j = 1, \ldots, m$, then the mean squared error of $\tilde{\theta}^{(1)}$ and $\tilde{\theta}^{(2)}$ satisfies

$$E\|(\tilde{\theta}^{(1)} - \theta_0)\|_2^2 \leq \frac{C_{13}}{\sum_{j=1}^m n_j \exp(\alpha_0)} + \frac{C_{14}}{\{n_{\min}\exp(\alpha_0)\}^4},$$

and

$$E\|(\tilde{\theta}^{(2)} - \theta_0)\|_2^2 \leq \frac{C_{15}}{\sum_{j=1}^m n_j \exp(\alpha_0)} + \frac{C_{16}}{\{n_1\exp(\alpha_0)\}^5},$$

respectively, where $n_{\min} = \min(n_1, \ldots, n_m)$ and $C_{13}, \ldots, C_{16}$ are constants unrelated to m, n_j and α_0 $(j = 1, \ldots, m)$.

B.8. Values for Simulation Studies

To mimic the rare outcomes or exposures, we varied the parameters p_1, p_2, α, and β.

Table B.1: Parameter Scenarios with Parameter Values.

Scenario	**Values**
common exposure common disease	$p_1 = 0.3$, $p_2 = 0.3$, $\alpha = -1$, $\beta = (-1, 1)$
rare exposure common diseases	$p_1 = 0.06$, $p_2 = 0.04$, $\alpha = -1$, $\beta = (-3, -3)$
common exposure rare disease	$p_1 = 0.2$, $p_2 = 0.7$, $\alpha = -3$, $\beta = (-4, -4)$
rare exposure rare disease	$p_1 = 0.04$, $p_2 = 0.06$, $\alpha = -4$, $\beta = (-4, -4)$

B.9. Estimation results for real-case study

We showed the estimation results for other risk factors in Figure B.1.

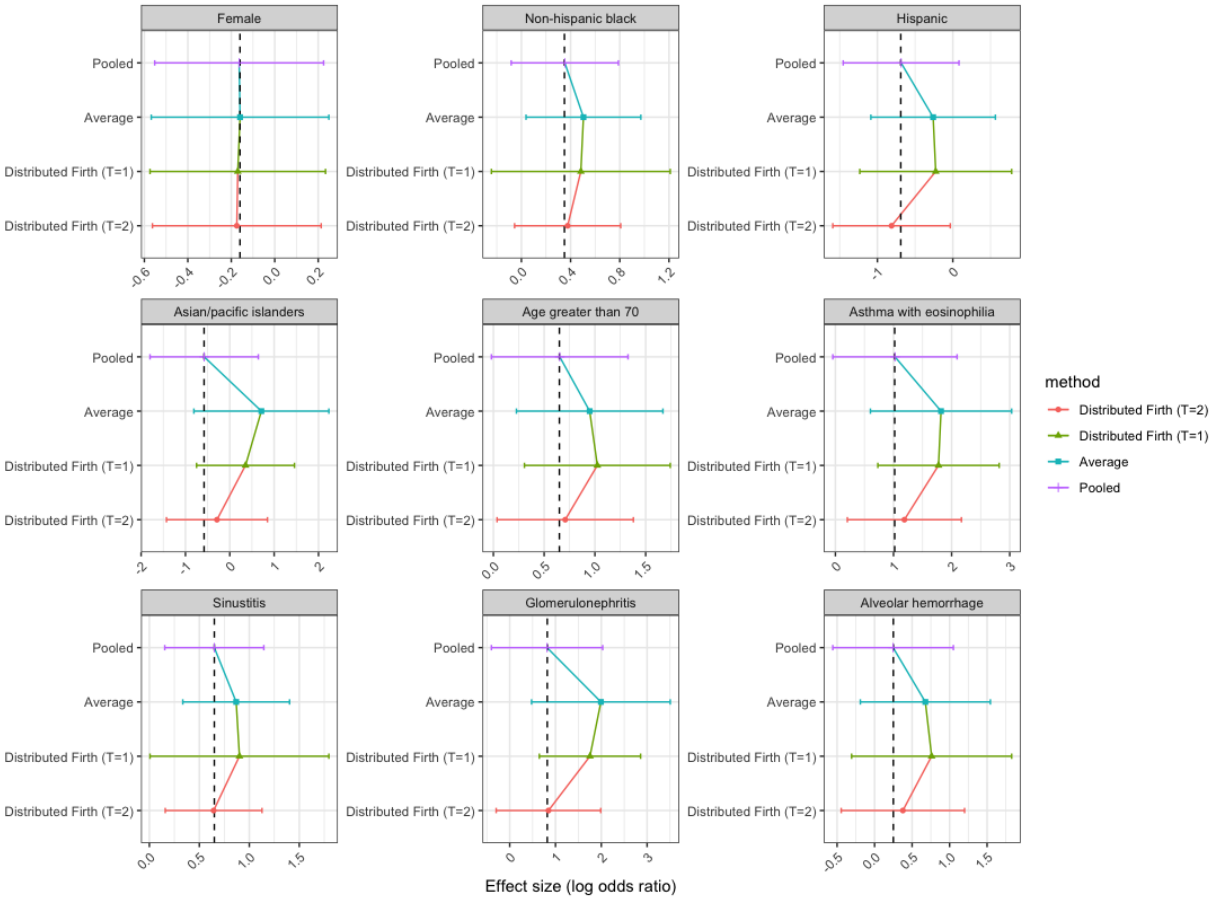

Figure B.1: Boxplot of estimates and 95 percentage confidence interval of the simple average estimator, distributed Firth estimators, and the pooled estimator.

APPENDIX C

SUPPLEMENTARY APPENDIX FOR CHAPTER 4

C.1. Sites in the PEDSnet Database

Participating institutions in the PEDSnet database included Cincinnati Children's Hospital Medical Center, Children's Hospital of Philadelphia, Children's Hospital of Colorado, Ann and Robert H. Lurie Children's Hospital of Chicago, Nationwide Children's Hospital, Nemours Children's Health System (in Delaware and Florida), Seattle Children's Hospital, and Stanford Children's Health.

C.2. PASC Symptoms and Conditions

This study included 24 PASC symptoms and conditions: abdominal pain, abnormal liver enzyme, acute kidney injury, acute respiratory distress syndrome, arrhythmias, cardiovascular signs and symptoms, changes in taste and smell, chest pain, cognitive functions, fatigue and malaise, fever and chills, fluid and electrolyte, generalized pain, hair loss, headache, heart disease, mental health disorders, musculoskeletal pain, myocarditis, myositis, Postural Orthostatic Tachycardia Syndrome (POTS) or dysautonomia, respiratory signs and symptoms, skin symptoms, and thrombophlebitis and thromboembolism.

The ICD and SNOMED diagnosis codes are available in Razzaghi et al. (2024). The classification of symptoms and conditions into systematic and syndromic groups are listed as follows: Systemic (diagnosed health conditions): acute kidney injury, acute respiratory distress syndrome, arrhythmias, fluid and electrolyte, heart disease, myocarditis, myositis, POTS or dysautonomia, and thrombophlebitis and thromboembolism. Syndromic (symptoms, signs, non-specific laboratory abnormalities): abdominal pain, abnormal liver enzyme, cardiovascular signs and symptoms, changes in taste and smell, chest pain, cognitive functions, fatigue and malaise, fever and chills, generalized pain, hair loss, headache, mental health disorders, musculoskeletal pain, respiratory signs and symptoms, skin symptoms.

C.3. Sensitivity Analysis of Different Sets of NCOs

The proposed method's robustness is affected by the selection of NCOs. Table C.1 showed the 62 NCOs that were manually reviewed by the clinicians. Originally, we selected 30 NCOs as the sensitivity analysis, which included acne, closed fracture of distal end of radius, closed injury of head, concussion, contact dermatitis, diaper rash, displacements - bone, falls, foreign body in ear, impetigo, inguinal hernia, injury of finger, injury of free lower limb, injury of head, injury of left leg, injury of right foot, injury of right hand, injury of right leg, injury of upper extremity, insect bite, plagiocephaly, scoliosis, snoring or obstructive sleep apnea, speech delay, speech dysfunction, sprain of ankle, tinea capitis, tinea corporis, tongue tie, umbilical hernia, and wax in ear or impacted cerumen.

Table C.1: The full list of 62 NCOs. The NCO list has been manually reviewed by medical clinicians.

NCO Name
Acanthosis nigricans
Acne
Alkalosis
Anemia neoplastic disease
Biliary calculus
Bronchiectasis
Burping
Cannabis abuse
Chalazion
Closed fracture of distal end of radius
Closed injury of head
Concussion
Contact dermatitis
Curvature spine
Cystic fibrosis

NCO Name
Diaper rash
Displacements - bone
Epilepsy
Esotropia
Falls
Foreign body in ear
Hashimoto thyroiditis
Hemochromatosis
Hyperlipidemia
Hypoparathyroidism
Impetigo
Inguinal hernia
Injury of finger
Injury of free lower limb
Injury of head
Injury of left leg
Injury of right foot
Injury of right hand
Injury of right leg
Injury of upper extremity
Insect bite
Jaundice
Leukemia
Low back pain
Lump right breast
Myopia
Nystagmus

NCO Name
Opioid dependence
Osteodystrophy
Perforation of tympanic membrane
Plagiocephaly
Postural kyphosis
Raynaud's disease
Scoliosis
Secondary malignant neoplasm large intestine
Snoring or obstructive sleep apnea
Solitary nodule of lung
Speech delay
Speech dysfunction
Sprain of ankle
Tinea capitis
Tinea corporis
Tinea pedis
Tongue tie
Umbilical hernia
Vitamin D deficiency
Wax in ear or impacted cerumen

Table C.2 showed the systematic bias by using full list of NCOs or partial list of NCOs. From this sensitivity analysis the estimation from the systematic bias does not change.

C.4. Simulation Setup And Performance in MSE

While our approach is a general method that could be used for both binary and Poisson outcomes, simulation studies are performed with scenarios generating relative risk as examples. Assume that

Table C.2: Sensitivity analysis for estimated systematic bias from the partial NCO list (30 NCOs) versus the full list (62 NCOs).

Minority Group	**Systematic Bias (Full NCO List)**	**Systematic Bias (Partial NCO List)**
AAPI	0.75	0.74
Hispanic	1.02	1.01
NHB	0.88	0.88

there are n patients in total. For the i-th patient, the observed confounding variables are denoted as $\mathbf{X}_i$, which is of p-dimension. For $\mathbf{X}_i$, the categorical variables were generated by a uniform distribution and cut into different categories with cutoff values. The binomial variables were generated with different probabilities of success.

The unmeasured confounders are denoted as $\mathbf{U}_i$, which is of q-dimension. For $\mathbf{U}_i$, the binomial variables were generated with different probabilities of success, and variables with a normal distribution were generated with mean zero and standard deviation of 2. An intercept term was included automatically in both $\mathbf{X}_i$ and $\mathbf{U}_i$. We set $p = 7$ and $q = 3$. The average treatment effect τ was set to -1. $M = 40$ NCOs are generated.

To evaluate type I error rates, we generated the systematic error of the m-th NCO via the normal distribution $\mathcal{N}(b, \sigma^2)$, where b was set to 0 and σ was set to 0.20. The function $\beta(A_i, \mathbf{X}_i)$ of measured confounders to the potential outcomes for the intervention and control groups was set to:

$$\beta(0, \mathbf{X}_i) = \mathbf{X}_i^\top(-1,\ 0.5,\ 0.3,\ 0.6,\ 0.5,\ -0.1,\ -0.5,\ 0.1,\ 0.5),$$
$$\beta(1, \mathbf{X}_i) = \mathbf{X}_i^\top(-2,\ 1,\ 0.6,\ 1.2,\ 0.8,\ -0.6,\ -1,\ 0.4,\ 1).$$

The function $\gamma_t(\mathbf{U}_i)$ of the time-varying effect of unmeasured confounders on the potential outcomes was set to:

$$\gamma_0(\mathbf{U}_i) = \gamma_1(\mathbf{U}_i) = \mathbf{U}_i^\top(-2,\ 0.5,\ 0.3,\ -1).$$

We ranged sample sizes as 1000, 2000, and 3000.

To evaluate statistical power, $b_0 \sim \mathcal{N}(b, \sigma^2)$ was ranged as 0.05, 0.10, and 0.15. We set:

$$\gamma_1(\mathbf{U}_i) = \mathbf{U}_i^\top(-2,\ 0.5,\ 0.3,\ -1) + b_0.$$

Other parameters were generated the same as in the previous paragraph. To show the estimated bias and coverage probabilities for the proposed calibration method compared to the DiD method, we set $b = 0.5$ and $\tau(\mathbf{X}_i) = -1$. The other parameters were generated the same as in the previous paragraph.

Table C.3 showed the mean square error for comparison of the proposed method with the baseline method before the calibration.

Table C.3: The mean square error (MSE) across different sample sizes. The MSE is defined as the average of squared differences between the estimated treatment effects and the true treatment effect across simulation replicates.

Method	**N = 1000**	**N = 2000**	**N = 3000**
Before calibration	0.425	0.352	0.331
After calibration	0.149	0.065	0.039

C.5. Details of Data Generation for Invalid NCOs

To evaluate the scenario that includes partially invalid NCOs, we generated the invalid NCOs as follows. Specifically, we assume that the first $M_1 = \frac{9}{10}M$ are valid NCOs generated as described previously. The remaining $M - M_1$ are invalid NCOs, denoted $W_{i,t,m}$, and are generated as:

$$\log\left[\mathbb{E}(W_{i,t,m} \mid A_i = 0, \mathbf{X}_i, \mathbf{U}_i)\right] = \beta(A_i, \mathbf{X}_i) + \gamma_t(\mathbf{U}_i),$$
$$\log\left[\mathbb{E}(W_{i,t,m} \mid A_i = 1, \mathbf{X}_i, \mathbf{U}_i)\right] = \beta(A_i, \mathbf{X}_i) + \xi(\mathbf{X}_i) + \gamma_1(\mathbf{U}_i).$$

Here, $\xi(\mathbf{X}_i) = 4b_m$, where $b_m \sim \mathcal{N}(4b, \sigma^2)$. We varied the sample sizes as 1000, 2000, and 3000. We set $b = 0.5$ and $\tau(\mathbf{X}_i) = -1$. All other parameters were generated as described in the previous paragraph.

C.6. Details about NC-DiD

DiD Method

A traditional difference-in-differences (DiD) model can be applied to estimate the target estimand. It requires the *parallel trends assumption*. On the logarithm scale—appropriate for the risk ratio estimand—the parallel trends assumption can be stated as:

$$\log\left(\frac{\mathbb{E}[Y_1(0) \mid A=1]}{\mathbb{E}[Y_1(0) \mid A=0]}\right) - \log\left(\frac{\mathbb{E}[Y_0(0) \mid A=1]}{\mathbb{E}[Y_0(0) \mid A=0]}\right) = 0. \tag{C.1}$$

Under this assumption, the following equation holds:

$$\mathbb{E}[Y_1(0) \mid A=1] = \left(\frac{\mathbb{E}[Y_0 \mid A=1]}{\mathbb{E}[Y_0 \mid A=0]}\right) \times \mathbb{E}[Y_1 \mid A=0]. \tag{C.2}$$

Assumption 1 (Consistency).

$$Y \mid (T=t) = Z_t Y_t(1) + (1 - Z_t) Y_t(0). \tag{C.3}$$

By Assumption 1, the conditional treatment effect on the treated group, expressed as a risk ratio, is identified by:

$$\tau = \left(\frac{\mathbb{E}[Y_1 \mid A=1]}{\mathbb{E}[Y_1 \mid A=0]}\right) \times \left(\frac{\mathbb{E}[Y_0 \mid A=0]}{\mathbb{E}[Y_0 \mid A=1]}\right). \tag{C.4}$$

We denote by $\tau^\star$ the right-hand side of the above equation, which corresponds to:

$$\tau^\star = \exp(\beta_3). \tag{C.5}$$

The quantity $\tau^\star$ is always identifiable from the observed data but may not be equal to the true effect τ if the parallel trends assumption does not hold. The effect of unmeasured confounders can be quantified by:

$$b_0 = \log\left(\frac{\mathbb{E}[Y_1(0) \mid A=1]}{\mathbb{E}[Y_1(0) \mid A=0]}\right) - \log\left(\frac{\mathbb{E}[Y_0(0) \mid A=1]}{\mathbb{E}[Y_0(0) \mid A=0]}\right) = \log(\tau) - \log(\tau^\star). \tag{C.6}$$

We refer to b_0 as the *unmeasured confounder bias* for the outcome of interest. Since the mean of the control counterfactual for the treated group is not generally identifiable, the target estimand τ is not identifiable either.

Negative Control Outcome

A *negative control outcome* (NCO) is a random variable that is not causally affected by the treatment but is influenced by unmeasured confounding variables related to the main outcome of interest. In other words, a statistical association between M NCOs and the treatment assignment would exist only if unmeasured confounders are present.

In the absence of unmeasured confounding, it follows that for the m-th NCO, the systematic bias $b_m = 0$ for $m = 1, \ldots, M$. Thus, a nonzero b_m signals a violation of the unconfoundedness assumption. Importantly, b_m is identifiable from the observed data, which is the foundation of our proposed hypothesis testing method.

Proposed Negative Control-Calibrated Method for DiD (NC-DiD)

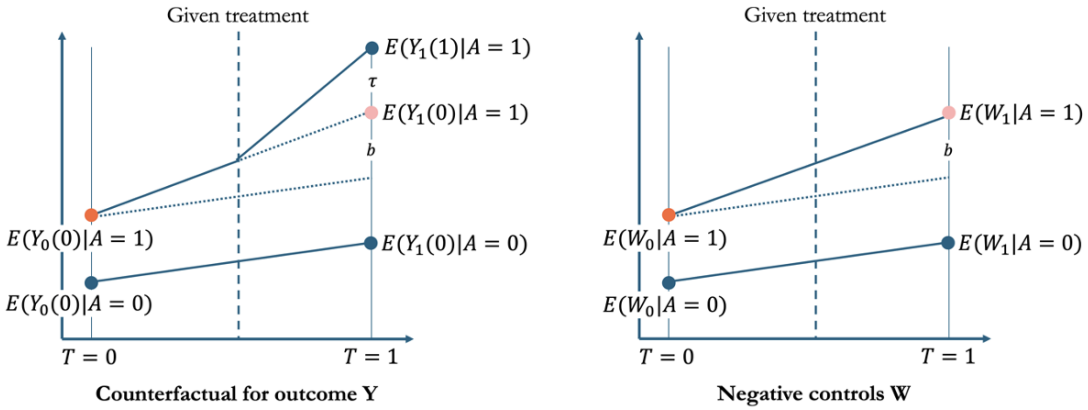

Figure C.1: Illustration of systematic bias. The left panel showed the notation for the counterfactual outcome Y changes over time. The right panel demonstrated the notation for the negative control outcome W.

Figure C.1 illustrates the proposed calibration method NC-DiD. The bias b_0 corrects for the biased conditional treatment effect among treated subjects. Specifically $\log(\tau) = \beta_3 - b_0$. If the bias b_0 can be identified from negative control outcomes, then the target estimand τ is also identifiable.

We consider the following problems:

- How can we perform hypothesis testing of the parallel trends assumption?
- How can we calibrate $\tau^\star$ by estimating the systematic bias b so that we can identify the target estimand τ?

We can solve the problem with the help of negative control outcomes (NCOs) at time $T = 0$ and $T = 1$. Suppose we have access to M NCOs. We denote the m-th NCO at times 0 and 1 as W_{m0} and W_{m1} for $m = 1, \ldots, M$. We define the NCO-revealed confounding bias for the m-th NCO as:

$$b_m = \log\left(\frac{\mathbb{E}[W_{m1} \mid A=1]}{\mathbb{E}[W_{m1} \mid A=0]}\right) - \log\left(\frac{\mathbb{E}[W_{m0} \mid A=1]}{\mathbb{E}[W_{m0} \mid A=0]}\right). \tag{C.7}$$

As discussed above, a nonzero NCO-revealed confounding bias is a signal of the presence of unmea-

sured confounding with time-varying effects. Note that unmeasured confounding for an NCO does not necessarily violate the parallel trends assumption for the outcome of interest. In other words, the unmeasured confounder U_t $(t = 0, 1)$ can cause heterogeneous effects on the association between treatment and NCOs and the outcome of interest.

To adjust for this heterogeneity and to identify $\tau^\star$, we posit Assumption 2 and Assumption 3 as follows:

Assumption 2 (NCO and systematic bias). *The NCO-revealed confounding biases* $\{b_m\}_{m=1}^M$ *are independent copies of a random variable* θ *with* $\theta \sim \mathcal{N}(b, \sigma^2)$*, where* b *represents the systematic bias and* σ^2 *the variance. The systematic bias from the outcome of interest also comes from* θ*.*

Assumption 2 is a distributional assumption, implying that the systematic bias affecting the outcome of interest follows a similar distributional pattern as the systematic bias observed in the negative control outcomes. In this way, we are able to identify the value of b_0 from the NCOs. The parameter $b = 0$ can be interpreted as a scenario where the parallel trends assumption essentially holds, and nonzero b_m values occur due to random error.

Assumption 3 (Outcome regression model). *The following underlying semi-parametric models hold:*

$$\log \mathbb{E}[Y_T \mid A, X, U_T] = \alpha_{0,0} + \alpha_{1,0}A + \alpha_{2,0}T + \alpha_{3,0}T \times A + \gamma_0(X) + \eta(U_T), \tag{C.8}$$

$$\log \mathbb{E}[W_{m,T} \mid A, X, U_T] = \alpha_{0,m} + \alpha_{1,m}A + \alpha_{2,m}T + \alpha_{3,m}T \times A + \gamma_m(X) + \eta_m(U_T), \tag{C.9}$$

for some square integrable functions $\gamma_m(X)$ *and constants* $\alpha_{i,m}$ *for* $m = 0, \ldots, M$ *and* $i = 0, 1, 2, 3$*. By the definition of NCO,* $\alpha_{3,m} = 0$*, and* $\alpha_{3,0} = \log(\tau)$*.*

Recall from the main manuscript, the following equations represent the working models on the matched cohort:

$$\log \mathbb{E}[Y_T \mid A, X] = \beta_0 + \beta_1 A + \beta_2 T + \beta_3 T \times A, \tag{C.10}$$

$$\log \mathbb{E}[W_{m,T} \mid A, X] = \beta_{0,m} + \beta_{1,m} A + \beta_{2,m} T + \beta_{3,m} T \times A. \tag{C.11}$$

It is worthwhile to note that under Assumption 3, we have $b_m = \beta_{3,m}$.

Estimation Procedure

With the help of Assumptions 1, 2, and 3, we can estimate the target estimand. We first conduct propensity score matching on X. A logistic regression model is fitted using the measured confounders, and treated individuals are matched to control individuals based on their estimated propensity scores. In the matched cohort, we estimate the treatment effect. Let $\hat{\beta}_3$ be the estimator for $\tau^\star$, and let $\hat{\sigma}^2_{\beta_3}$ be its estimated asymptotic variance.

To calibrate the estimate of $\tau^\star$, we fit the model for each m-th NCO ($m = 1, \ldots, M$). Let $\hat{b}_m$ denote the estimator of $\beta_{3,m} = b_m$, and let $\hat{\sigma}^2_m$ be its asymptotic variance.

From Assumption 2, we have:

$$\hat{b}_m \mid b_m \sim \mathcal{N}(b_m, \hat{\sigma}^2_m), \quad \text{and} \quad b_m, b_0 \sim \mathcal{N}(b, \sigma^2).$$

We then estimate b and σ^2 by maximizing the likelihood after integrating b_m out. As a byproduct of estimating b, we perform a hypothesis test for the parallel trends assumption using the null hypothesis $H_0 : b = 0$. The test statistic is:

$$S = \frac{|\hat{b}|}{\sqrt{V}}, \quad \text{where} \quad V = \left(\sum_{m=1}^{M} \frac{1}{\hat{\sigma}^2_m + \hat{\sigma}^2} \right)^{-1}.$$

The two-sided p-value for the hypothesis is given by:

$$\text{p-value} = 2 \times \{1 - \Phi(|S|)\},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Finally, we obtain the calibrated estimate $\hat{\tau}$ using the fact that, asymptotically:

$$\hat{\beta}_3 \sim \mathcal{N}(\tau + \hat{b}, \hat{\sigma}^2_{\beta_3} + \hat{\sigma}^2).$$

The calibrated estimator is:

$$\hat{\tau} = \exp(\hat{\beta}_3 - \hat{b}).$$

The 95% confidence interval for τ is given by:

$$\left(\exp\left(\hat{\beta}_3 - \hat{b} - 1.96\sqrt{\hat{\sigma}^2_{\beta_3} + \hat{\sigma}^2}\right), \quad \exp\left(\hat{\beta}_3 - \hat{b} + 1.96\sqrt{\hat{\sigma}^2_{\beta_3} + \hat{\sigma}^2}\right)\right).$$

C.7. Theoretical Guarantee of NC-DiD

We provide the theoretical guarantees of our proposed method as follows. Proposition 1 considers the validity for the hypothesis testing with respect to the parallel trends assumption. Proposition 2 characterizes the asymptotic behavior of the calibrated estimator. This is used to construct a valid confidence interval for τ.

Proposition 1 (Asymptotic Normality for Bias Estimation). *Suppose that estimators b_m are independent, and $n \to \infty$ while M is fixed. Then $\hat{b}$ and $\hat{\sigma}^2$ converge in probability to the true values $b^\star$ and $\sigma^{\star 2}$, respectively; that is,*

$$\hat{b} \xrightarrow{p} b^\star \quad \text{and} \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^{\star 2}.$$

Moreover, we have

$$\sqrt{MV}(\hat{b} - b^\star) \xrightarrow{d} \mathcal{N}(0, 1).$$

Standard asymptotic theory yields that the maximum likelihood estimate from the model shown in Equation (2) satisfies

$$\Omega^{1/2}(\hat{\beta} - \beta^\star) \xrightarrow{d} \mathcal{N}(0, I_4),$$

where I_4 denotes the 4-dimensional identity matrix, and

$$\Omega = \left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(\hat{\beta}; Y_i, Z_i)\right) \left[\frac{1}{n}\sum_{i=1}^{n} \nabla f(\hat{\beta}; Y_i, Z_i) \nabla f(\hat{\beta}; Y_i, Z_i)^\top\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(\hat{\beta}; Y_i, Z_i)\right),$$

where f is the loss function of the working model. Let $\kappa = (\Omega^{-1})_{4,4}$ be the fourth diagonal entry of Ω^{-1}.

Proposition 2 (Asymptotic Normality for NC-DiD Estimator). *Under the assumptions imposed in Lemma 1 and some regularity conditions, we have*

$$\sqrt{1/\kappa}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, 1).$$

APPENDIX D

SUPPLEMENTARY APPENDIX FOR CHAPTER 5

D.1. Description of electronic health records (EHR) data

The real-world data utilized in our analysis is derived from electronic health records (EHRs), covering a wide range of healthcare interaction information routinely collected and stored by hospitals. This includes clinical data such as diagnoses and treatments, laboratory and test results, and administrative data including patient demographics and billing information. The hospital-based EHR data from the Researching COVID to Enhance Recovery (RECOVER) Initiative COVID-19 Database served as the basis for defining and determining exposure, outcomes, and covariates. Unlike General Practitioner (GP) data, self-reported data, or external data sources, our study used the structured, standardized EHR entries made by healthcare providers within hospital settings. EHR data provides a more detailed and integrated view of a patient's health status, medical history, and healthcare interactions across various providers and settings.

RECOVER population and generalizability

The National Institutions of Health (NIH) launched the new RECOVER initiative in 2021 to leverage electronic health record (EHR) data to better identify and characterize patients with post-acute sequelae of SARS-CoV-2 infection (PASC). RECOVER obtains EHRs from three large national healthcare networks within the United States, covering regional catchment areas across 41 states. These networks collectively hold the EHRs of over 60 million patients, including records from more than 7 million individuals who have been affected by COVID-19. RECOVER collaborates with the National Institutes of Health's (NIH) All of Us Research Program, which contributes additional health records to this vast database. Together, these sources comprise one of the world's largest collections of EHRs.

List of negative control outcomes

Table D.1: The 31 Negative control outcomes used in the analysis.

No.	Health Condition
1	Acne
2	Closed fracture of distal end of radius
3	Closed injury of head
4	Concussion
5	Contact dermatitis
6	Diaper rash
7	Displacements - bone
8	Falls
9	Foreign body in ear
10	Impetigo
11	Inguinal hernia
12	Injury of finger
13	Injury of free lower limb
14	Injury of head
15	Injury of left leg
16	Injury of right foot
17	Injury of right hand
18	Injury of right leg
19	Injury of upper extremity
20	Insect bite
21	Plagiocephaly
22	Scoliosis
23	Snoring/Obstructive sleep apnea
24	Speech delay
25	Speech dysfunction
26	Sprain of ankle
27	Tinea capitis
28	Tinea corporis
29	Tongue tie
30	Umbilical hernia
31	Wax in ear/impacted cerumen

Study Variables

Table D.2: Confounding Variables Used in the Propensity Model and Other Variables for Eligibility Criteria

Variables	Functional form	Values	Detail	Codes/References
Age (years)	Linear	Numerical	Records in person domain	Age = integer(date–birth date)/365.25
Sex	Indicator	Male/Female	Records in person domain	NA
Obesity	Indicator	Yes/No	Measurement domain; Age < 24×30.5 days: NHANES weight z > 1.64 Age 24×30.5 to 240×30.5 days: NHANES BMI z > 1.64 Age ≥ 240×30.5 days: BMI > 30 kg/m^2	NA
PMCA	3 categories	0: No chronic condition 1: Non-complex chronic 2: Complex chronic	Condition & visit domains	Simon et al. (2018)
Severity of acute COVID-19	4 categories	Asymptomatic, Mild, Moderate, Severe	Condition & visit domains	Forrest et al. (2022)
Number of ED visits	4 categories	0,1,2,≥3	Condition & visit domains	NA
Number of inpatient visits	4 categories	0,1,2,≥3	Inpatient stay, ED admit inpatient, Observation Stay	NA
Number of outpatient visits	4 categories	0,1,2,≥3	Ambulatory visit, Interactive telemedicine	NA
Number of unique medications	4 categories	0,1,2,≥3	Drug exposure domain	RxNorm code
Chronic condition clusters	Indicator	Yes/No	205 chronic clusters (condition/visit domains)	Rao et al. (2022)
Other eligibility criteria variables				
Prior encounter	Indicator	Yes/No	Condition & visit domains	NA
Documented SARS-CoV-2	Indicator	Yes/No	PCR, serology, antigen test, or COVID-19/PASC diagnosis	https://github.com/PEDSnet/PASC
Exposure variable				
Race/Ethnicity	4 categories	NHW, NHB, Hispanic, AAPI	race_concept_id, ethnicity_concept_id	https://github.com/PEDSnet/Data_Models_Public
Outcome variable				
PASC symptoms/conditions	24 categories	Abdominal pain, fatigue, myocarditis, cognitive dysfunction, etc.	Condition/visit domains; presence of relevant ICD codes	Razzaghi et al. (2024)

Note: All domains in the table above are based on the PEDSnet common data model (CDM). More details are available through this link: https://data-models-service.research.chop.edu.

Classification of Symptoms and Conditions into Systematic and Syndromic Groups

Table D.3: Classification of Symptoms and Conditions into Systematic and Syndromic Groups

Syndromic (symptoms, signs, non-specific laboratory abnormalities)	**Systemic (diagnosed health conditions)**
• Abdominal pain • Abnormal liver enzyme • Cardiovascular signs and symptoms • Changes in taste and smell • Chest pain • Cognitive functions • Fatigue and malaise • Fever and chills • Generalized pain • Hair loss • Headache • Mental health disorders • Musculoskeletal pain • Respiratory signs and symptoms • Skin symptoms	• Acute kidney injury • Acute respiratory distress syndrome • Arrhythmias • Fluid and electrolyte • Heart disease • Myocarditis • Myositis • Postural Orthostatic Tachycardia Syndrome (POTS) or dysautonomia • Thrombophlebitis and thromboembolism

The distribution of COVID-19-positive patients by hospital and cohort entry time

Table D.4: The distribution of COVID-19 positive patients by hospital and cohort entry time

	Severe[1]				Non-severe[2]			
Hospital	**NHW**	**AAPI**	**Hispanic**	**NHB**	**NHW**	**AAPI**	**Hispanic**	**NHB**
A	1459 (16.0%)	37 (6.3%)	154 (4.1%)	389 (11.6%)	14323 (14.3%)	530 (5.2%)	1226 (2.2%)	6069 (14.3%)
B	1113 (12.2%)	95 (16.1%)	317 (8.4%)	646 (19.3%)	24713 (24.7%)	1960 (19.1%)	4289 (7.6%)	9949 (23.4%)
C	1031 (11.3%)	57 (9.7%)	801 (21.2%)	163 (4.9%)	4415 (4.4%)	242 (2.4%)	3880 (6.9%)	848 (2.0%)
D	390 (4.3%)	29 (4.9%)	159 (4.2%)	257 (7.7%)	4647 (4.7%)	471 (4.6%)	2245 (4.0%)	4443 (10.5%)
E	420 (4.6%)	13 (2.2%)	60 (1.6%)	48 (1.4%)	7581 (7.6%)	322 (3.1%)	1094 (1.9%)	1047 (2.5%)
F	712 (7.8%)	71 (12.0%)	595 (15.7%)	250 (7.5%)	1907 (1.9%)	239 (2.3%)	1833 (3.3%)	661 (1.6%)
G	754 (8.2%)	28 (4.7%)	92 (2.4%)	153 (4.6%)	7547 (7.6%)	742 (7.2%)	768 (1.4%)	1314 (3.1%)
H	232 (2.5%)	1 (0.2%)	16 (0.4%)	35 (1.0%)	4529 (4.5%)	109 (1.1%)	283 (0.5%)	710 (1.7%)
I	1156 (12.6%)	85 (14.4%)	183 (4.8%)	517 (15.5%)	8530 (8.5%)	1098 (10.7%)	1801 (3.2%)	5844 (13.8%)
J	882 (9.6%)	33 (5.6%)	643 (17.0%)	501 (15.0%)	8173 (8.2%)	526 (5.1%)	5824 (10.4%)	3404 (8.0%)
K	105 (1.1%)	25 (4.2%)	273 (7.2%)	64 (1.9%)	6783 (6.8%)	2568 (25.0%)	28574 (50.8%)	5440 (12.8%)
L	182 (2.0%)	88 (14.9%)	320 (8.5%)	100 (3.0%)	1942 (1.9%)	1151 (11.2%)	3170 (5.6%)	1057 (2.5%)
M	704 (7.7%)	28 (4.7%)	173 (4.6%)	223 (6.7%)	4792 (4.8%)	318 (3.1%)	1239 (2.2%)	1691 (4.0%)
Entry time								
03/2020 - 06/2020	136 (1.5%)	17 (2.9%)	163 (4.3%)	122 (3.6%)	907 (0.9%)	190 (1.8%)	1935 (3.4%)	891 (2.1%)
07/2020 - 10/2020	241 (2.6%)	19 (3.2%)	166 (4.4%)	108 (3.2%)	5454 (5.5%)	433 (4.2%)	4967 (8.8%)	1760 (4.1%)
11/2020 - 02/2021	753 (8.2%)	47 (8.0%)	346 (9.1%)	291 (8.7%)	16011 (16.0%)	1240 (12.1%)	10089 (17.9%)	5262 (12.4%)
03/2021 - 06/2021	720 (7.9%)	32 (5.4%)	277 (7.3%)	309 (9.2%)	7916 (7.9%)	445 (4.3%)	3007 (5.3%)	4056 (9.5%)
07/2021 - 10/2021	1063 (11.6%)	37 (6.3%)	394 (10.4%)	485 (14.5%)	13868 (13.9%)	818 (8.0%)	6095 (10.8%)	6477 (15.2%)
11/2021 - 02/2022	3050 (33.4%)	193 (32.7%)	1320 (34.9%)	1136 (34.0%)	33830 (33.9%)	3746 (36.5%)	19028 (33.8%)	16290 (38.4%)
03/2022 - 06/2022	1662 (18.2%)	120 (20.3%)	577 (15.2%)	447 (13.4%)	12030 (12.0%)	1951 (19.0%)	4939 (8.8%)	3739 (8.8%)
07/2022 - 10/2022	1515 (16.6%)	125 (21.2%)	543 (14.3%)	448 (13.4%)	9866 (9.9%)	1453 (14.1%)	6166 (11.0%)	4002 (9.4%)

D.2. Supplemental Results: crude analysis comparing COVID-19 positive and negative patients

Definition of COVID-19-negative group

For COVID-19-negative patients, we included patients who had neither a documented SARS-CoV-2 infection nor a diagnosis of multisystem inflammatory syndrome in children (MIS-C) within the same study period, and who had at least one negative COVID-19 test result. A random negative test was chosen as the index date for COVID-19-negative patients. The selection of participants for COVID-19-negative in real-world data is summarized in Figure S13. The COVID-19-negative patients were required to be members of the healthcare organization during the previous 18 months and have at least one visit 28 to 179 days after the index date, as specified in the flowchart. These included subjects come from ED visits or other visits (i.e., routine, physical), primary care, office visits, and others.

For COVID-19-negative patients, we determined severity status based on the symptoms and conditions observed during the acute phase (seven days before to thirteen days after the index date).

Primary care settings typically provide a more comprehensive representation of the general population due to their broader patient demographics and the nature of routine and preventive care they offer. Thus, the presence of these major primary care-driven systems in our data source bolsters the external validity and generalizability of our results to the broader child and adolescent population.

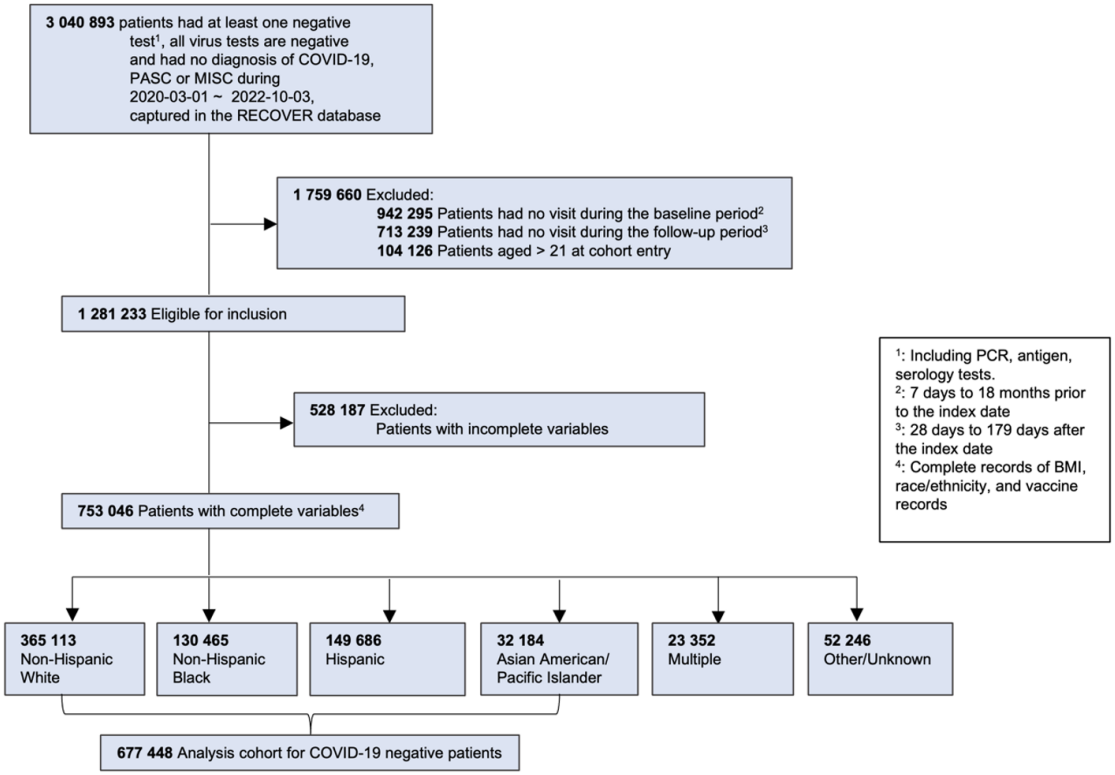

Figure D.1: Selection of participants for COVID-19-negative patients.

Baseline characteristics of COVID-19-negative group

Table D.5: Baseline characteristics of COVID-19 negative patients, by race/ethnicity and severity status. Abbreviation: NHW: Non-Hispanic White; NHB: Non-Hispanic Black; AAPI: Asian American or Pacific Islander.

	Severe				Non-severe			
Characteristic	**NHW**	**AAPI**	**Hispanic**	**NHB**	**NHW**	**AAPI**	**Hispanic**	**NHB**
N	87222	5709	26637	24024	277891	26475	123049	106441
Age <1	10895 (12.5%)	711 (12.5%)	3461 (13.0%)	3978 (16.6%)	24652 (8.9%)	2313 (8.7%)	11795 (9.6%)	10748 (10.1%)
Age 1–<5	24363 (27.9%)	1692 (29.6%)	8024 (30.1%)	7646 (31.8%)	75138 (27.0%)	7739 (29.2%)	32990 (26.8%)	31213 (29.3%)
Age 5–<12	23450 (26.9%)	1727 (30.3%)	7382 (27.7%)	5815 (24.2%)	82081 (29.5%)	8493 (32.1%)	41893 (34.0%)	33954 (31.9%)
Age 12–<16	14085 (16.1%)	771 (13.5%)	4087 (15.3%)	3175 (13.2%)	45195 (16.3%)	3128 (11.8%)	19255 (15.6%)	15311 (14.4%)
Age 16–<21	14429 (16.5%)	808 (14.2%)	3683 (13.8%)	3410 (14.2%)	50825 (18.3%)	4802 (18.1%)	17116 (13.9%)	15215 (14.3%)
Sex Female	38451 (44.1%)	2374 (41.6%)	11735 (44.1%)	10568 (44.0%)	138595 (49.9%)	13015 (49.2%)	60993 (49.6%)	52848 (49.7%)
Sex Male	48771 (55.9%)	3335 (58.4%)	14902 (55.9%)	13456 (56.0%)	139296 (50.1%)	13460 (50.8%)	62056 (50.4%)	53593 (50.4%)
Obesity = 0	59812 (68.6%)	4299 (75.3%)	16878 (63.4%)	14958 (62.3%)	190501 (68.6%)	17846 (67.4%)	54986 (44.7%)	59931 (56.3%)
Obesity = 1	27410 (31.4%)	1410 (24.7%)	9759 (36.6%)	9066 (37.7%)	87390 (31.4%)	8629 (32.6%)	68063 (55.3%)	46510 (43.7%)
PMCA = 0	48327 (55.4%)	3241 (56.8%)	15290 (57.4%)	13804 (57.5%)	186253 (67.0%)	19569 (73.9%)	91266 (74.2%)	71779 (67.4%)
PMCA = 1	16614 (19.0%)	1029 (18.0%)	4902 (18.4%)	4602 (19.2%)	51283 (18.5%)	3952 (14.9%)	18525 (15.1%)	20753 (19.5%)
PMCA = 2	22281 (25.5%)	1439 (25.2%)	6445 (24.2%)	5618 (23.4%)	40355 (14.5%)	2954 (11.2%)	13258 (10.8%)	13909 (13.1%)
Vaccine = 0	79029 (90.6%)	4866 (85.2%)	23943 (89.9%)	22208 (92.4%)	243699 (87.7%)	21074 (79.6%)	105725 (85.9%)	96004 (90.2%)
Vaccine = 1	1387 (1.6%)	149 (2.6%)	512 (1.9%)	401 (1.7%)	5670 (2.0%)	923 (3.5%)	3101 (2.5%)	2242 (2.1%)
Vaccine ≥ 2	6806 (7.8%)	694 (12.2%)	2182 (8.2%)	1415 (5.9%)	28522 (10.3%)	4478 (16.9%)	14223 (11.6%)	8195 (7.7%)

For COVID-19-negative patients, we determined severity status based on the symptoms and conditions observed during the acute phase (seven days before to fourteen days after the index date). The index date refers to the first visit with a COVID-19 negative test.

Raw incidence of PASC symptoms and conditions We calculated the incidence of PASC symptoms and conditions in both COVID-19 positive and negative cohorts stratified by severity. For each PASC symptom or condition, we calculated its incidence by dividing the number of patients who experienced the symptom or condition during the follow-up period but not at baseline by the total number of patients.

Rerun DiD method with NCO calibration on the COVID-19 negative cohort

In this sensitivity analysis, we rerun the DiD method with NCO calibration on the COVID-19 negative cohort. The results are presented in Figure D.2.

Table D.6: Comparison of post-acute conditions between COVID-19 positive and negative individuals

	All			Severe			Non-severe		
	Pos (%) (N=225,723)	Neg (%) (N=677,448)	p-value[1]	Pos (%) (N=16,862)	Neg (%) (N=143,592)	p-value[2]	Pos (%) (N=208,861)	Neg (%) (N=533,856)	p-value[3]
At least one condition	26.86	21.44	1.23e-279	34.17	19.80	1.56e-141	26.42	21.84	2.23e-180
Systematic conditions	2.69	1.87	6.33e-113	7.46	2.93	7.41e-166	2.37	1.61	2.05e-101
Acute kidney injury	0.22	0.15	1.35e-09	1.32	0.42	9.38e-52	0.13	0.08	1.13e-08
Acute respiratory distress syndrome	0.03	0.01	1.01e-07	0.24	0.03	2.77e-26	0.01	0.01	0.01
Arrhythmias	1.43	0.95	9.93e-78	4.41	1.56	2.45e-131	1.21	0.79	1.86e-63
Fluid and electrolyte	0.56	0.39	2.32e-28	3.33	0.92	3.20e-152	0.36	0.24	2.92e-17
Heart disease	0.37	0.28	2.44e-10	1.67	0.62	4.16e-49	0.27	0.20	5.11e-10
Myocarditis	0.03	0.01	9.89e-18	0.17	0.02	6.21e-21	0.02	0.002	1.72e-13
Myositis	0.02	0.02	0.01	0.08	0.02	8.39e-05	0.02	0.01	0.07
POTS/dysautonomia	1.03	0.71	1.65e-49	1.34	0.72	5.00e-17	1.01	0.71	8.49e-39
Thrombophlebitis and thromboembolism	0.13	0.11	1.79e-03	1.02	0.33	1.05e-40	0.06	0.05	0.01
Syndromic conditions	26.15	20.83	2.18e-282	32.48	18.77	4.17e-144	25.75	21.35	3.16e-173
Abdominal pain	3.04	2.27	8.55e-85	3.66	2.15	1.70e-31	2.99	2.31	4.47e-60
Abnormal liver enzyme	0.32	0.24	1.11e-09	1.17	0.43	6.44e-36	0.25	0.19	3.87e-07
Cardiovascular signs and symptoms	1.16	0.88	4.62e-31	1.69	0.86	5.37e-25	1.12	0.89	2.37e-19
Changes in taste and smell	0.16	0.04	3.49e-69	0.08	0.02	1.62e-05	0.17	0.05	3.45e-55
Chest pain	1.39	0.74	2.27e-173	1.69	0.68	5.66e-43	1.37	0.75	4.97e-133
Cognitive functions	0.68	0.66	0.35	1.00	0.62	1.09e-08	0.66	0.68	0.38
Fatigue and malaise	1.74	1.30	1.45e-50	3.22	1.74	7.60e-37	1.63	1.18	1.37e-49
Fever and chills	5.67	3.73	9.37e-306	9.15	3.31	6.57e-246	5.41	3.84	3.98e-177
Generalized pain	1.26	0.91	8.94e-47	1.78	1.02	2.67e-18	1.22	0.88	5.78e-40
Hair loss	0.24	0.14	1.50e-24	0.48	0.14	8.68e-23	0.22	0.14	1.36e-15
Headache	2.20	1.51	3.02e-99	2.33	1.23	2.45e-29	2.19	1.59	5.25e-65
Mental health	6.32	5.45	1.72e-44	8.04	5.05	2.96e-46	6.19	5.55	1.03e-21
Musculoskeletal pain	3.53	2.79	1.32e-66	4.03	2.72	8.21e-20	3.50	2.81	1.63e-50
Respiratory signs and symptoms	9.68	7.25	2.25e-236	15.02	6.77	4.65e-221	9.31	7.38	6.20e-134
Skin symptoms	3.62	2.70	1.04e-101	4.99	2.66	1.26e-56	3.51	2.71	6.47e-70

[1]P-value for all: Two-sided Z-test comparing COVID-19 positive and negative cohorts.
[2]P-value for severe: Two-sided Z-test comparing COVID-19 positive and negative cohorts among patients with severe cases.
[3]P-value for non-severe: Two-sided Z-test comparing COVID-19 positive and negative cohorts among patients with non-severe cases.

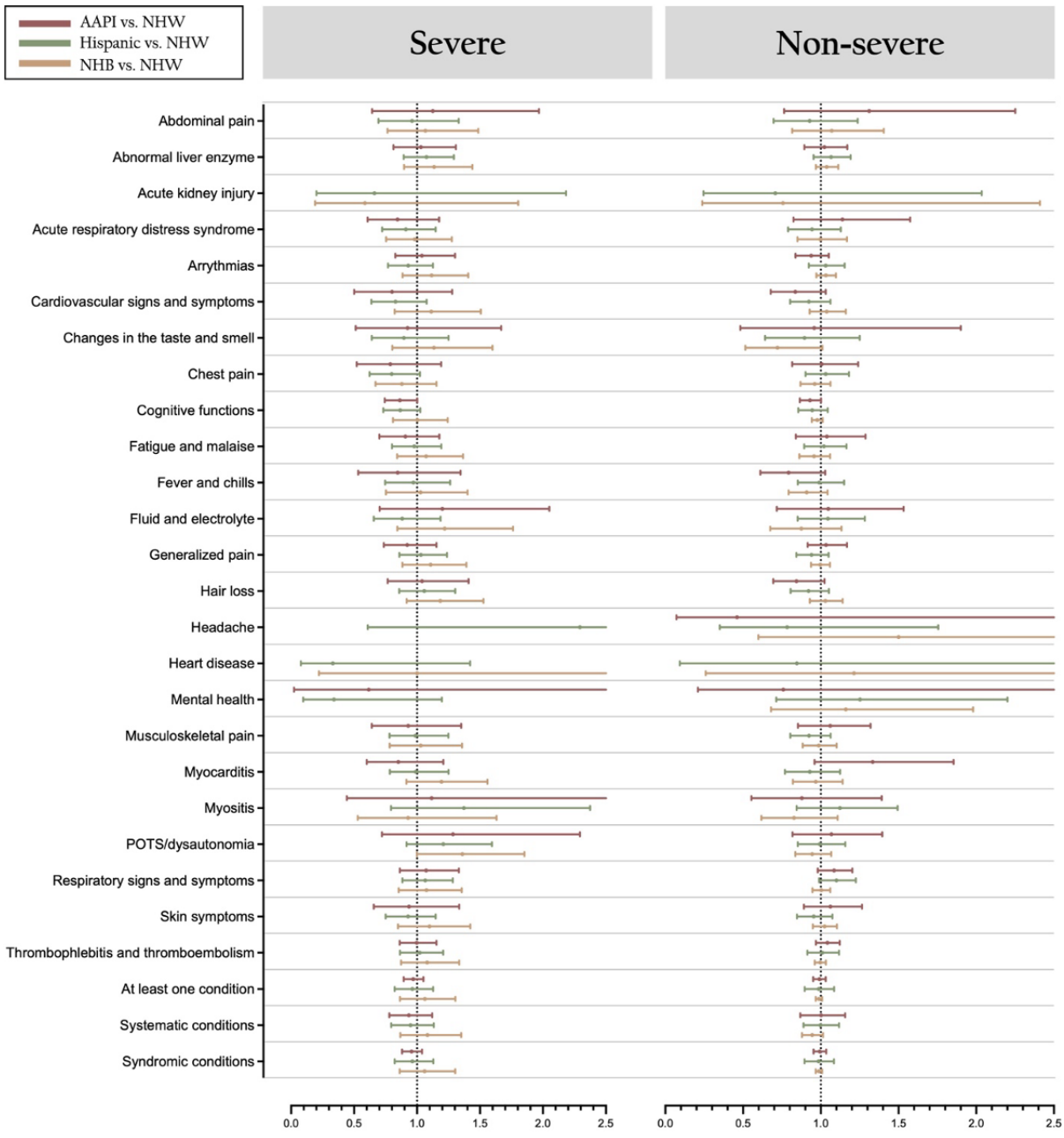

Figure D.2: The racial/ethnic difference in risk ratio (RR) across PASC symptoms and conditions for COVID-19 negative cohort (n=677,448 patients). The dots in the figure show the estimated RRs for each minority group versus the NHW group from the difference-in-differences (DiD) analysis with NCO calibration stratified by severity status. The error bars showed the 95% confidence interval (CI) of the estimated RR.

D.3. Supplemental Results: sensitivity analysis using standard regression models

We conducted a sensitivity analysis using standard regression models, adjusting for demographic, clinical, and healthcare utilization factors. Figure D.3 shows the results of this sensitivity analysis.

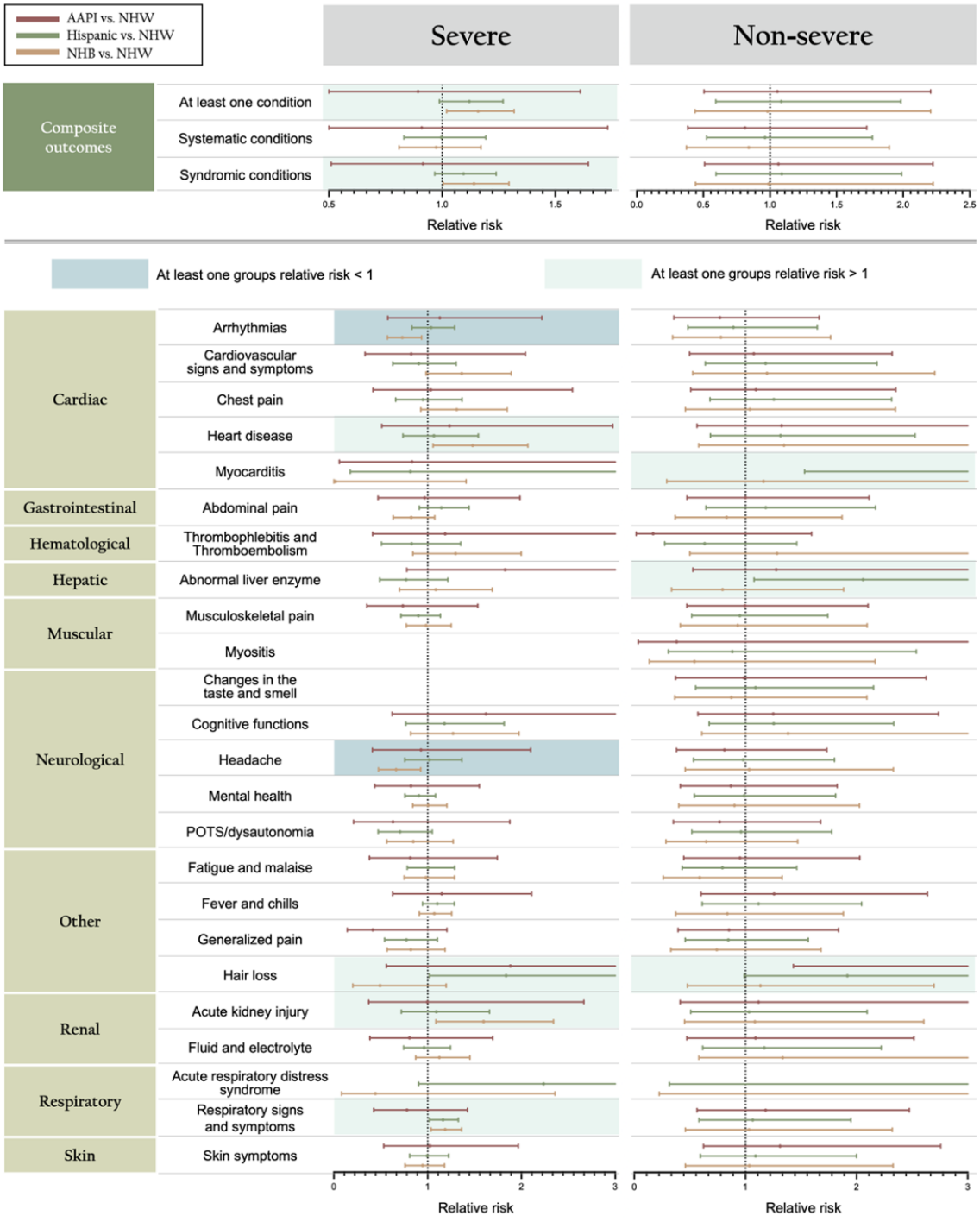

Figure D.3: Racial/Ethnic Differences that are attributable to COVID-19 estimated from the difference-in-differences (DiD) analyses for prevalence of PASC symptoms and conditions among COVID-19-positive patients (n=225,723 participants) using standard regression models, by race/ethnicity and severity status. The dots in the figure show the estimated RRs for each minority group versus the NHW group from the standard Poisson regression model stratified by severity status. The error bars showed the 95% confidence interval (CI) of the estimated RR.

APPENDIX E

SUPPLEMENTARY APPENDIX FOR CHAPTER 6

Table E.1 listed the used NCOs for the calibration of the digital twins.

Table E.1: The list of the negative control outcomes used for investigation in SPRINT-MIND and iSTAGING data.

Category	**Examples**
Frontal Lobe	Right FO (frontal operculum), Left FO (frontal operculum), Right MFC (medial frontal cortex), Left MFC (medial frontal cortex), Right MFG (middle frontal gyrus), Left MFG (middle frontal gyrus), Right MSFG (superior frontal gyrus medial segment), Left MSFG (superior frontal gyrus medial segment), Right SFG (superior frontal gyrus), Left SFG (superior frontal gyrus)
Insular Lobe	Right CO (central operculum), Left CO (central operculum)
Limbic System	Left PCgG (posterior cingulate gyrus), Right PCgG (posterior cingulate gyrus)
Occipital Lobe	Left Cun (cuneus), Right Cun (cuneus), Left LiG (lingual gyrus), Right LiG (lingual gyrus), Right IOG (inferior occipital gyrus), Right MOG (middle occipital gyrus), Left MOG (middle occipital gyrus), Right OCP (occipital pole), Left OCP (occipital pole), Right SOG (superior occipital gyrus), Left IOG (inferior occipital gyrus), Left SOG (superior occipital gyrus), Right Calc (calcarine cortex), Left Calc (calcarine cortex)
Parietal Lobe	Right PCu (precuneus), Left PCu (precuneus), Right PO (parietal operculum), Left PO (parietal operculum), Left PoG (postcentral gyrus), Right PoG (postcentral gyrus)
Temporal Lobe	Right PP (planum polare), Left PP (planum polare), Right FuG (fusiform gyrus), Left FuG (fusiform gyrus), Left ITG (inferior temporal gyrus), Right MTG (middle temporal gyrus), Left MTG (middle temporal gyrus), Right PT (planum temporale), Left PT (planum temporale), Right ITG (inferior temporal gyrus), Right PHG (parahippocampal gyrus), Left PHG (parahippocampal gyrus)

BIBLIOGRAPHY

Alberto Abadie. Semiparametric difference-in-differences estimators. *The review of economic studies*, 72(1):1–19, 2005.

Amy P Abernethy, James Gippetti, Rohit Parulkar, and Cindy Revol. Use of electronic health record data for quality reporting. *Journal of oncology practice*, 13(8):530–534, 2017.

Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34:61–88, 2011.

Joe Amoah, Elizabeth A Stuart, Sara E Cosgrove, Anthony D Harris, Jennifer H Han, Ebbing Lautenbach, and Pranita D Tamma. Comparing propensity score methods versus traditional regression analysis for the evaluation of observational data: a case study evaluating the treatment of gram-negative bloodstream infections. *Clinical Infectious Diseases*, 71(9):e497–e505, 2020.

Heather Angier, David Ezekiel-Herrera, Miguel Marino, Megan Hoopes, Elizabeth A Jacobs, Jennifer E DeVoe, and Nathalie Huguet. Racial/ethnic disparities in health insurance and differences in visit type for a population of patients with diabetes after medicaid expansion. *Journal of health care for the poor and underserved*, 30(1):116–130, 2019.

David A Asch, Md Nazmul Islam, Natalie E Sheils, Yong Chen, Jalpa A Doshi, John Buresh, and Rachel M Werner. Patient and hospital factors associated with differences in mortality rates among black and white us medicare beneficiaries hospitalized with covid-19 infection. *JAMA network open*, 4(6):e2112842–e2112842, 2021a.

David A Asch, Natalie E Sheils, Md Nazmul Islam, Yong Chen, Rachel M Werner, John Buresh, and Jalpa A Doshi. Variation in us hospital mortality rates for patients admitted with covid-19 during the first 6 months of the pandemic. *JAMA internal medicine*, 181(4):471–478, 2021b.

Peter C Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009.

Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

Sindhura Bandi, Michael Z Nevid, and Mahboobeh Mahdavinia. African american children are at higher risk of covid-19 infection. *Pediatric Allergy and Immunology*, 31(7):861, 2020.

Cristina Bezzio, Gionata Fiorino, Davide G Ribaldone, Alessandro Armuzzi, and Simone Saibeni. Ibd flare in the covid-19 pandemic: Therapy discontinuation is to blame. *Inflammatory Bowel Diseases*, 29(5):834–836, 2023.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John wiley & sons, 2021.

Barry C Burden, David T Canon, Kenneth R Mayer, and Donald P Moynihan. Election laws, mobilization, and turnout: The unanticipated consequences of election reform. *American Journal of Political Science*, 58(1):95–109, 2014.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230, 2021.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *Annals of Statistics*, 46:1541–1572, 2018.

Jiajie Chen, Yuqing Lei, Qiong Wu, Ting Zhou, Bingyu Zhang, Michael J Becich, Yuriy Bisyuk, Saul Blecker, Elizabeth A Chrischilles, Dimitri A Christakis, et al. Vaccine effectiveness among 5-to 17-year-old individuals with prior sars-cov-2 infection: An ehr-based target trial emulation study from the recover project. *medRxiv*, pages 2025–02, 2025.

Yong Chen, Jing Ning, and Chunyan Cai. Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics*, 16(4):727–739, 2015.

Yong Chen, Jianqiao Wang, Jessica Chubak, and Rebecca A Hubbard. Inflation of type i error rates due to differential misclassification in ehr-derived outcomes: empirical illustration using breast cancer recurrence. *Pharmacoepidemiology and drug safety*, 28(2):264–268, 2019.

Deena J Chisolm, Ryan Webb, Katherine S Salamon, Julia Schuchard, Eneida A Mendonca, Marion R Sills, Payal B Patel, Jordan Musante, Christopher B Forrest, Ravi Jhaveri, et al. Associations between area-level health-related social factor indices and risk of acute covid-19: An ehr-based cohort study from the recover program. *medRxiv*, pages 2022–12, 2022.

Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106:1–9, 2017.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972.

Hannah E Davis, Gina S Assaf, Lisa McCorkell, Hannah Wei, Ryan J Low, Yochai Re'em, Signe Redfield, Jared P Austin, and Athena Akrami. Characterizing long covid in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine*, 38, 2021.

Hannah E Davis, Lisa McCorkell, Julia Moore Vogel, and Eric J Topol. Long covid: major findings, mechanisms and recommendations. *Nature Reviews Microbiology*, 21(3):133–146, 2023.

Justin B Dimick and Andrew M Ryan. Methods for evaluating changes in health care policy: the

difference-in-differences approach. *Jama*, 312(22):2401–2402, 2014.

Rui Duan, Ming Cao, Yonghui Wu, Jing Huang, Joshua C Denny, Hua Xu, and Yong Chen. An empirical study for impacts of measurement errors on ehr based association studies. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1764, 2017.

Rui Duan, Mary Regina Boland, Jason H Moore, and Yong Chen. Odal: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 30–41. World Scientific, 2018.

Rui Duan, Mary Regina Boland, Zixuan Liu, Yue Liu, Howard H Chang, Hua Xu, Haitao Chu, Christopher H Schmid, Christopher B Forrest, John H Holmes, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385, 2020a.

Rui Duan, Chongliang Luo, Martijn J Schuemie, Jiayi Tong, C Jason Liang, Howard H Chang, Mary Regina Boland, Jiang Bian, Hua Xu, John H Holmes, et al. Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association*, 27(7):1028–1036, 2020b.

Rui Duan, Yang Ning, and Yong Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83, 2022.

Mackenzie J Edmondson, Chongliang Luo, Rui Duan, Mitchell Maltenfort, Zhaoyi Chen, Kenneth Locke Jr, Justine Shults, Jiang Bian, Patrick B Ryan, Christopher B Forrest, et al. An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes. *Scientific reports*, 11(1):19647, 2021.

Mackenzie J Edmondson, Chongliang Luo, Md Nazmul Islam, Natalie E Sheils, John Buresh, Zhaoyi Chen, Jiang Bian, and Yong Chen. Distributed quasi-poisson regression algorithm for modeling multi-site count outcomes in distributed data networks. *Journal of biomedical informatics*, 131: 104097, 2022.

Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

Chandra L Ford and Nina T Harawa. A new conceptualization of ethnicity for social epidemiologic and health equity research. *Social science & medicine*, 71(2):251–258, 2010.

Marvella E Ford and P Adam Kelly. Conceptualizing and categorizing race and ethnicity in health services research. *Health services research*, 40(5p2):1658–1675, 2005.

Christopher B Forrest, Peter A Margolis, L Charles Bailey, Keith Marsolo, Mark A Del Beccaro, Jonathan A Finkelstein, David E Milov, Veronica J Vieland, Bryan A Wolf, Feliciano B Yu, et al. Pedsnet: a national pediatric learning health system. *Journal of the American Medical Informatics Association*, 21(4):602–606, 2014.

Christopher B Forrest, Evanette K Burrows, Asuncion Mejias, Hanieh Razzaghi, Dimitri Christakis, Ravi Jhaveri, Grace M Lee, Nathan M Pajor, Suchitra Rao, Deepika Thacker, et al. Severity of acute covid-19 in children< 18 years old march 2020 to december 2021. *Pediatrics*, 149(4): e2021055765, 2022.

Charles P Friedman, Adam K Wong, and David Blumenthal. Achieving a nationwide learning health system. *Science translational medicine*, 2(57):57cm29–57cm29, 2010.

Landon Gibson and Frederick Zimmerman. Measuring the sensitivity of difference-in-difference estimates to the parallel trends assumption. *Research Methods in Medicine & Health Sciences*, 2 (4):148–156, 2021.

Mugdha Gokhale, Christopher F Bell, Scott Doyle, Jolyon Fairburn-Beech, Jonathan Steinfeld, and Melissa K Van Dyke. Prevalence of eosinophilic granulomatosis with polyangiitis and associated health care utilization among patients with concomitant asthma in us commercial claims database. *Journal of Clinical Rheumatology*, 27(3):107, 2021.

Sander Greenland, Judea Pearl, and James M Robins. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

Destin Groff, Ashley Sun, Anna E Ssentongo, Djibril M Ba, Nicholas Parsons, Govinda R Poudel, Alain Lekoubou, John S Oh, Jessica E Ericson, Paddy Ssentongo, et al. Short-term and long-term rates of postacute sequelae of sars-cov-2 infection: a systematic review. *JAMA network open*, 4 (10):e2128568–e2128568, 2021.

Dae Woong Ham and Luke Miratrix. Benefits and costs of matching prior to a difference in differences analysis when parallel trends does not hold. *The Annals of Applied Statistics*, 18(3): 2096–2122, 2024.

Sebastien Haneuse and Kyu Ha Lee. Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circulation: Cardiovascular Quality and Outcomes*, 9(3):322–331, 2016.

Sune Welling Hansen. Polity size and local political trust: A quasi-experiment using municipal mergers in d enmark. *Scandinavian Political Studies*, 36(1):43–66, 2013.

Georg Heinze and Michael Schemper. A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, 57(1):114–119, 2001.

Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression.

Statistics in Medicine, 21:2409–2419, 2002.

William R Hogan, Elizabeth A Shenkman, Temple Robinson, Olveen Carasquillo, Patricia S Robinson, Rebecca Z Essner, Jiang Bian, Gigi Lipori, Christopher Harle, Tanja Magoc, et al. The OneFlorida Data Trust: a centralized, translational research data infrastructure of statewide scope. *Journal of the American Medical Informatics Association*, 29(4):686–693, 2022.

Chanelle J Howe, Stephen R Cole, Bryan Lau, Sonia Napravnik, and Joseph J Eron Jr. Selection bias due to loss to follow up in cohort studies. *Epidemiology*, 27(1):91–97, 2016.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. In *MEDINFO 2015: eHealth-enabled Health*, pages 574–578. IOS Press, 2015.

Kosuke Imai, In Song Kim, and Erik H Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605, 2023.

Rupert W Jakes, Namhee Kwon, Beth Nordstrom, Rebecca Goulding, Kyle Fahrbach, Jialu Tarpey, and Melissa K Van Dyke. Burden of illness associated with eosinophilic granulomatosis with polyangiitis: a systematic literature review and meta-analysis. *Clinical Rheumatology*, pages 1–8, 2021.

Naimin Jing, Yiwen Lu, Jiayi Tong, James Weaver, Patrick Ryan, Hua Xu, and Yong Chen. Evaluating the bias, type i error and statistical power of the prior knowledge-guided integrated likelihood estimation (pie) for bias reduction in ehr based association studies. *Journal of Biomedical Informatics*, page 104787, 2025.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2019.

Ashwini Joshi, Angelika Geroldinger, Lena Jiricka, Pralay Senchaudhuri, Christopher Corcoran, and Georg Heinze. Solutions to problems of nonexistence of parameter estimates and sparse data bias in poisson regression. *Statistical Methods in Medical Research*, 31(2):253–266, 2022.

Evangelia Katsoulakis, Qi Wang, Huanmei Wu, Leili Shahriyari, Richard Fletcher, Jinwei Liu, Luke Achenie, Hongfang Liu, Pamela Jackson, Ying Xiao, et al. Digital twins for health: a scoping review. *NPJ digital medicine*, 7(1):77, 2024.

Dhruv Khullar, Yongkang Zhang, Chengxi Zang, Zhenxing Xu, Fei Wang, Mark G Weiner, Thomas W Carton, Russell L Rothman, Jason P Block, and Rainu Kaushal. Racial/ethnic disparities in post-acute sequelae of sars-cov-2 infection in new york: an ehr-based cohort study

from the recover program. *Journal of General Internal Medicine*, 38(5):1127–1136, 2023.

Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2): 137–163, 2001.

Michal Kori, Maya Avidan, and Chani Topf-Olivestone. Long-term follow-up and predictors of complicated disease behavior in pediatric crohn's disease patients. *Journal of Pediatric Gastroenterology and Nutrition*, 74(4):471–475, 2022.

Ioannis Kosmidis and David Firth. A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, 4:1097–1112, 2010.

Ioannis Kosmidis and David Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108:71–82, 2021.

Subra Kugathasan, Lee A Denson, Thomas D Walters, Mi-Ok Kim, Urko M Marigorta, Melanie Schirmer, Kajari Mondal, Chunyan Liu, Anne Griffiths, Joshua D Noe, et al. Prediction of complicated disease course for children newly diagnosed with crohn's disease: a multicentre inception cohort study. *The Lancet*, 389(10080):1710–1718, 2017.

Eileen T Lake, Bingyu Zhang, Jiayi Tong, Gloria Mpundu, Kelsey N Gross, Malini H Correa, Lauren B Wethman, Lynne Moronski, Domenique Villani, and Yong Chen. International benchmarks for the practice environment scale of the nursing work index: A meta-analysis. *Nursing & Health Sciences*, 26(4):e70013, 2024.

Ellen H Lee, Kelsey L Kepler, Anita Geevarughese, Rachel Paneth-Pollak, Marie S Dorsinville, Stephanie Ngai, and Kathleen H Reilly. Race/ethnicity among children with covid-19–associated multisystem inflammatory syndrome. *JAMA Network Open*, 3(11):e2030280, 2020.

Arie Levine, Anne Griffiths, James Markowitz, David C Wilson, Dan Turner, Richard K Russell, John Fell, Frank M Ruemmele, Thomas Walters, Mary Sherlock, et al. Pediatric modification of the montreal classification for inflammatory bowel disease: the paris classification. *Inflammatory bowel diseases*, 17(6):1314–1321, 2011.

Lu Li, Ting Zhou, Yiwen Lu, Jiajie Chen, Yuqing Lei, Qiong Wu, Jonathan Arnold, Michael J Becich, Yuriy Bisyuk, Saul Blecker, et al. Kidney function following covid-19 in children and adolescents. *JAMA network open*, 8(4):e254129–e254129, 2025.

Ruowang Li, Joseph D Romano, Yong Chen, and Jason H Moore. Centralized and federated models for the analysis of clinical data. *Annual review of biomedical data science*, 7, 2024.

Dan-Yu Lin and Daniel Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 2010.

Xiaokang Liu, Rui Duan, Chongliang Luo, Alexis Ogdie, Jason H Moore, Henry R Kranzler, Jiang

Bian, and Yong Chen. Multisite learning of high-dimensional heterogeneous data with applications to opioid use disorder study of 15,000 patients across 5 clinical sites. *Scientific reports*, 12(1): 11073, 2022.

Yu-Lun Liu, Bingyu Zhang, Haitao Chu, and Yong Chen. Network meta-analysis made simple: A composite likelihood approach. *Research Synthesis Methods*, page 1–19, 2025. doi: 10.1017/rsm.2024.12.

Sandra Lopez-Leon, Talia Wegman-Ostrosky, Carol Perelman, Rosalinda Sepulveda, Paulina A Rebolledo, Angelica Cuapio, and Sonia Villapol. More than 50 long-term effects of covid-19: a systematic review and meta-analysis. *Scientific reports*, 11(1):16144, 2021.

Sandra Lopez-Leon, Talia Wegman-Ostrosky, Norma Cipatli Ayuzo del Valle, Carol Perelman, Rosalinda Sepulveda, Paulina A Rebolledo, Angelica Cuapio, and Sonia Villapol. Long-covid in children and adolescents: a systematic review and meta-analyses. *Scientific reports*, 12(1):9950, 2022.

Yiwen Lu, Jiayi Tong, Jessica Chubak, Thomas Lumley, Rebecca A Hubbard, Hua Xu, and Yong Chen. Leveraging error-prone algorithm-derived phenotypes: Enhancing association studies for risk factors in ehr data. *Journal of Biomedical Informatics*, 157:104690, 2024.

Yiwen Lu, Jiayi Tong, Dazheng Zhang, Jiajie Chen, Lu Li, Yuqing Lei, Ting Zhou, Leyna V Aragon, Michael J Becich, Saul Blecker, et al. Does sars-cov-2 infection increase risk of neuropsychiatric and related conditions? findings from difference-in-differences analyses. *Research Square*, 2025.

Chongliang Luo, Rui Duan, Adam C Naj, Henry R Kranzler, Jiang Bian, and Yong Chen. Odach: a one-shot distributed algorithm for cox model with heterogeneous multi-center data. *Scientific reports*, 12(1):6627, 2022a.

Chongliang Luo, Md Nazmul Islam, Natalie E Sheils, John Buresh, Martijn J Schuemie, Jalpa A Doshi, Rachel M Werner, David A Asch, and Yong Chen. dpql: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling. *Journal of the American Medical Informatics Association*, 29(8):1366–1371, 2022b.

Edmund J Malesky, Cuong Viet Nguyen, and Anh Tran. The impact of recentralization on public services: A difference-in-differences analysis of the abolition of elected councils in vietnam. *American Political Science Review*, 108(1):144–168, 2014.

Yoav Mazor, Itay Maza, Eduard Kaufman, Shomron Ben-Horin, Amir Karban, Yehuda Chowers, and Rami Eliakim. Prediction of disease complication occurrence in crohn's disease using phenotype and genotype parameters at diagnosis. *Journal of Crohn's and Colitis*, 5(6):592–597, 2011.

Peter McCullagh. *Tensor methods in statistics*. Courier Dover Publications, 2018.

Rosa Morello, Francesco Mariani, Luca Mastrantoni, Cristina De Rose, Giuseppe Zampino, Daniel

Munblit, Louise Sigfrid, Piero Valentini, and Danilo Buonsenso. Risk factors for post-covid-19 condition (long covid) in children: a prospective cohort study. *EClinicalMedicine*, 59, 2023.

Stephen Morris, Rachael M Hunter, Angus IG Ramsay, Ruth Boaden, Christopher McKevitt, Catherine Perry, Nanik Pursani, Anthony G Rudd, Lee H Schwamm, Simon J Turner, et al. Impact of centralising acute stroke services in english metropolitan areas on mortality and length of hospital stay: difference-in-differences analysis. *Bmj*, 349, 2014.

Md Nasir, Nicole Cook, Daniel Parras, Sumit Mukherjee, Geralyn Miller, Juan Lavista Ferres, and Katherine Chung-Bridges. Using data science and a health equity lens to identify long-covid sequelae among medically underserved populations. *Journal of Health Care for the Poor and Underserved*, 34(2):521–534, 2023.

Tahmina Nasserie, Michael Hittle, and Steven N Goodman. Assessment of the frequency and variety of persistent symptoms among patients with covid-19: a systematic review. *JAMA network open*, 4(5):e2111417–e2111417, 2021.

National Center for Health Statistics, U.S. Census Bureau, and Household Pulse Survey. Long covid. https://www.cdc.gov/nchs/covid19/pulse/long-covid.htm, 2022. Accessed: 2025-03-16.

Michael Neureiter. Evaluating the effects of immigrant integration policies in western europe using a difference-in-differences approach. *Journal of Ethnic and Migration Studies*, 45(15):2779–2800, 2019.

Lauren L O'Mahoney, Ash Routen, Clare Gillies, Winifred Ekezie, Anneka Welford, Alexa Zhang, Urvi Karamchandani, Nikita Simms-Williams, Shabana Cassambai, Ashkon Ardavani, et al. The prevalence and long-term health effects of long covid among hospitalised and non-hospitalised populations: A systematic review and meta-analysis. *EClinicalMedicine*, 55, 2023.

World Health Organization et al. A clinical case definition for post covid-19 condition in children and adolescents by expert consensus, 16 february 2023. In *A clinical case definition for post COVID-19 condition in children and adolescents by expert consensus, 16 February 2023*. World Health Organization, 2023.

Art B Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(4), 2007.

Laura Pazzagli, Marie Linder, Mingliang Zhang, Emese Vago, Paul Stang, David Myers, Morten Andersen, and Shahram Bahmanyar. Methods for time-varying exposure related problems in pharmacoepidemiology: an overview. *Pharmacoepidemiology and drug safety*, 27(2):148–160, 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Suchitra Rao, Grace M Lee, Hanieh Razzaghi, Vitaly Lorman, Asuncion Mejias, Nathan M Pajor, Deepika Thacker, Ryan Webb, Kimberley Dickinson, L Charles Bailey, et al. Clinical features and

burden of postacute sequelae of sars-cov-2 infection in children and adolescents. *JAMA pediatrics*, 176(10):1000–1009, 2022.

Hanieh Razzaghi, Christopher B Forrest, Kathryn Hirabayashi, Qiong Wu, Andrea J Allen, Suchitra Rao, Yong Chen, H Timothy Bunnell, Elizabeth A Chrischilles, Lindsay G Cowell, et al. Vaccine effectiveness against long covid in children. *Pediatrics*, 153(4):e2023064446, 2024.

David B Richardson, Ting Ye, and Eric J Tchetgen Tchetgen. Generalized difference-in-differences. *Epidemiology*, 34(2):167–174, 2023.

Alfonso J Rodriguez-Morales, María Camila Lopez-Echeverri, Maria Fernanda Perez-Raga, Valentina Quintero-Romero, Valentina Valencia-Gallego, Nicolas Galindo-Herrera, Santiago López-Alzate, Juan Diego Sánchez-Vinasco, Juan José Gutiérrez-Vargas, Percy Mayta-Tristan, et al. The global challenges of the long covid-19 in adults and children. *Travel Medicine and Infectious Disease*, 54:102606, 2023.

Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

Defne Saatci, Tom A Ranger, Cesar Garriga, Ash Kieran Clift, Francesco Zaccardi, Pui San Tan, Martina Patone, Carol Coupland, Anthony Harnden, Simon J Griffin, et al. Association between race and covid-19 outcomes among 2.6 million children in england. *JAMA pediatrics*, 175(9): 928–938, 2021.

Heinz Schmidli, Dieter A Häring, Marius Thomas, Adrian Cassidy, Sebastian Weber, and Frank Bretz. Beyond randomized clinical trials: use of external controls. *Clinical Pharmacology & Therapeutics*, 107(4):806–816, 2020.

Martijn J Schuemie, Patrick B Ryan, William DuMouchel, Marc A Suchard, and David Madigan. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in medicine*, 33(2):209–218, 2014.

Martijn J Schuemie, George Hripcsak, Patrick B Ryan, David Madigan, and Marc A Suchard. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11):2571–2577, 2018.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, Charles Fisher, Alzheimer's Disease Neuroimaging Initiative, et al. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2022.

Joel Selvakumar, Lise Beier Havdal, Martin Drevvatne, Elias Myrstad Brodwall, Lise Lund Berven, Tonje Stiansen-Sonerud, Gunnar Einvik, Truls Michael Leegaard, Trygve Tjade, Annika E Michelsen, et al. Prevalence and characteristics associated with post–covid-19 condition among

nonhospitalized adolescents and young adults. *JAMA Network Open*, 6(3):e235763–e235763, 2023.

Steven Shea and George Hripcsak. Accelerating the use of electronic health records in physician practices. *New England Journal of Medicine*, 362:192, 2010.

Yishan Shen, Bingyu Zhang, Ravi Jhaveri, Jiayi Tong, Ting Zhou, Jiajie Chen, Dazheng Zhang, Qiong Wu, Chengxi Zang, Fei Wang, et al. Assessment of nirmatrelvir-ritonavir effects on acute and post-acute covid-19 illness in us adolescents: Target trial emulation. *medRxiv*, pages 2025–04, 2025.

Rachel E Sherman, Steven A Anderson, Gerald J Dal Pan, Gerry W Gray, Thomas Gross, Nina L Hunter, Lisa LaVange, Danica Marinac-Dabic, Peter W Marks, Melissa A Robb, et al. Real-world evidence—what is it and what can it tell us. *N Engl J Med*, 375(23):2293–2297, 2016.

Di Shu, Kazuki Yoshida, Bruce H Fireman, and Sengwee Toh. Inverse probability weighted cox model in multi-site studies without sharing individual-level data. *Statistical methods in medical research*, 29(6):1668–1681, 2020.

Tamara D Simon, Wren Haaland, Katherine Hawley, Karen Lambka, and Rita Mangione-Smith. Development and validation of the pediatric medical complexity algorithm (pmca) version 3.0. *Academic pediatrics*, 18(5):577–580, 2018.

STAR Clinical Research Network. Star clinical research network. https://starcrn.org/, n.d. Accessed: April 17, 2023.

Adam J Streeter, Nan Xuan Lin, Louise Crathorne, Marcela Haasova, Christopher Hyde, David Melzer, and William E Henley. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *Journal of clinical epidemiology*, 87:23–34, 2017.

Elizabeth A Stuart, Haiden A Huskamp, Kenneth Duckworth, Jeffrey Simmons, Zirui Song, Michael E Chernew, and Colleen L Barry. Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Services and Outcomes Research Methodology*, 14:166–182, 2014.

Olivia V Swann, Karl A Holden, Lance Turtle, Louisa Pollock, Cameron J Fairfield, Thomas M Drake, Sohan Seth, Conor Egan, Hayley E Hardwick, Sophie Halpin, et al. Clinical characteristics of children and young people admitted to hospital with covid-19 in united kingdom: prospective multicentre observational cohort study. *bmj*, 370, 2020.

Huilin Tang, Bingyu Zhang, Yiwen Lu, William T Donahoo, Naykky Singh Ospina, Pareeta Kotecha, Ying Lu, Jiayi Tong, Steven M Smith, Eric I Rosenberg, et al. Assessing the benefit–risk profile of newer glucose-lowering drugs: A systematic review and network meta-analysis of randomized outcome trials. *Diabetes, Obesity and Metabolism*, 27(3):1444–1455, 2025.

Tanayott Thaweethai, Sarah E Jolley, Elizabeth W Karlson, Emily B Levitan, Bruce Levy, Grace A

McComsey, Lisa McCorkell, Girish N Nadkarni, Sairam Parthasarathy, Upinder Singh, et al. Development of a definition of postacute sequelae of sars-cov-2 infection. *Jama*, 329(22):1934–1946, 2023.

Jiayi Tong, Rui Duan, Ruowang Li, Martijn J Scheuemie, Jason H Moore, and Yong Chen. Robust-odal: Learning from heterogeneous health systems without sharing patient-level data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pages 695–706. World Scientific, 2019.

Jiayi Tong, Chongliang Luo, Md Nazmul Islam, Natalie E Sheils, John Buresh, Mackenzie Edmondson, Peter A Merkel, Ebbing Lautenbach, Rui Duan, and Yong Chen. Distributed learning for heterogeneous clinical data with application to integrating covid-19 data across 230 sites. *NPJ digital medicine*, 5(1):76, 2022.

Jiayi Tong, Yishan Shen, Alice Xu, Xing He, Chongliang Luo, Mackenzie Edmondson, Dazheng Zhang, Yiwen Lu, Chao Yan, Ruowang Li, et al. Evaluating site-of-care-related racial disparities in kidney graft failure using a novel federated learning framework. *Journal of the American Medical Informatics Association*, 31(6):1303–1312, 2024.

Helen Twohig, Ram Bajpai, Nadia Corp, Alice Faux-Nightingale, Christian Mallen, Toni Robinson, Glenys Somayajula, Danielle Van der Windt, Victoria Welsh, and Claire Burton. Long-term outcomes of covid-19 infection in children and young people: a systematic review and meta-analysis. *medRxiv*, pages 2023–04, 2023.

U.S. Food and Drug Administration. Real-world data: Assessing electronic health records and medical claims to support regulatory decision-making for drug and biological products: Guidance for industry, draft guidance. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory, September 2021. Accessed: April 25, 2025.

Kaushik P Venkatesh, Marium M Raza, and Joseph C Kvedar. Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation. *NPJ digital medicine*, 5(1):150, 2022.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014.

HaiYing Wang. More efficient estimation for logistic regression with optimal subsamples. *Journal of machine learning research*, 20, 2019.

HaiYing Wang. Logistic regression for massive data with rare events. In *International Conference*

on Machine Learning, pages 9829–9836. PMLR, 2020.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pages 3636–3645. PMLR, 2017.

Xuefeng Wang. Firth logistic regression for rare variant association tests. *Frontiers in Genetics*, 5: 187, 2014.

Yumeng Wang and Ziwei Zhu. Reboot: Distributed statistical learning via refitting bootstrap samples. *arXiv preprint arXiv:2207.09098*, 2022.

Michael E Wechsler, Praveen Akuthota, David Jayne, Paneez Khoury, Amy Klion, Carol A Langford, Peter A Merkel, Frank Moosig, Ulrich Specks, Maria C Cid, et al. Mepolizumab or placebo for eosinophilic granulomatosis with polyangiitis. *New England Journal of Medicine*, 376(20): 1921–1932, 2017.

Brian W Whitcomb and Ashley I Naimi. Defining, quantifying, and interpreting "noncollapsibility" in epidemiologic studies of measures of "effect". *American Journal of Epidemiology*, 190(5):697–700, 2021.

William Whittaker, Laura Anselmi, Søren Rud Kristensen, Yiu-Shing Lau, Simon Bailey, Peter Bower, Katherine Checkland, Rebecca Elvey, Katy Rothwell, Jonathan Stokes, et al. Associations between extending access to primary care and emergency department visits: a difference-in-differences analysis. *PLoS medicine*, 13(9):e1002113, 2016.

Coady Wing, Kosali Simon, and Ricardo A Bello-Gomez. Designing difference in difference studies: best practices for public health policy research. *Annual review of public health*, 39(1):453–469, 2018.

Qiong Wu, Martijn J Schuemie, Marc A Suchard, Patrick Ryan, George M Hripcsak, Charles A Rohde, and Yong Chen. Padé approximant meets federated learning: A nearly lossless, one-shot algorithm for evidence synthesis in distributed research networks with rare outcomes. *Journal of biomedical informatics*, 145:104476, 2023.

Qiong Wu, Nathan M Pajor, Yiwen Lu, Charles J Wolock, Jiayi Tong, Vitaly Lorman, Kevin B Johnson, Jason H Moore, Christopher B Forrest, David A Asch, et al. A latent transfer learning method for estimating hospital-specific post-acute healthcare demands following sars-cov-2 infection. *Patterns*, 5(11), 2024a.

Qiong Wu, Jiayi Tong, Bingyu Zhang, Dazheng Zhang, Jiajie Chen, Yuqing Lei, Yiwen Lu, Yudong Wang, Lu Li, Yishan Shen, et al. Real-world effectiveness of bnt162b2 against infection and severe diseases in children and adolescents. *Annals of internal medicine*, 177(2):165–176, 2024b.

Qiong Wu, Bingyu Zhang, Jiayi Tong, L Charles Bailey, H Timothy Bunnell, Jiajie Chen, Elizabeth A Chrischilles, Dimitri A Christakis, Stephen M Downs, Kathryn Hirabayashi, et al. Real-

world effectiveness and causal mediation study of bnt162b2 on long covid risks in children and adolescents. *EClinicalMedicine*, 79, 2025.

Ting Ye, Ashkan Ertefaie, James Flory, Sean Hennessy, and Dylan S Small. Instrumented difference-in-differences. *Biometrics*, 79(2):569–581, 2023.

Bingyu Zhang, Deepika Thacker, Ting Zhou, Dazheng Zhang, Yuqing Lei, Jiajie Chen, Elizabeth A Chrischilles, Dimitri A Christakis, Soledad Fernandez, Vidu Garg, et al. Cardiovascular post-acute sequelae of sars-cov-2 in children and adolescents: cohort study using electronic health records. *Nature communications*, 16(1):3445, 2025a.

Bingyu Zhang, Qiong Wu, Ravi Jhaveri, Ting Zhou, Michael J Becich, Yuriy Bisyuk, Frank Blanceró, Elizabeth A Chrischilles, Cynthia H Chuang, Linday G Cowell, et al. Reinfection with sars-cov-2 in the omicron era is associated with increased risk of post-acute sequelae of sars-cov-2 infection: A recover-ehr cohort study. *medRxiv*, 2025b.

Dazheng Zhang, Jiayi Tong, Naimin Jing, Yuchen Yang, Chongliang Luo, Yiwen Lu, Dimitri A Christakis, Diana Güthe, Mady Hornig, Kelly J Kelleher, et al. Learning competing risks across multiple hospitals: one-shot distributed algorithms. *Journal of the American Medical Informatics Association*, 31(5):1102–1112, 2024.

Dazheng Zhang, Ronen Stein, Yiwen Lu, Ting Zhou, Yuqing Lei, Lu Li, Jiajie Chen, Jonathan Arnold, Michael J Becich, Elizabeth A Chrischilles, et al. Pediatric gastrointestinal tract outcomes during the postacute phase of covid-19. *JAMA Network Open*, 8(2):e2458366–e2458366, 2025c.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. *Advances in Neural Information Processing Systems*, 25, 2012.

Bingqing Zhou, Jason Fine, Aurelien Latouche, and Myriam Labopin. Competing risks regression for clustered data. *Biostatistics*, 13(3):371–383, 2012.

Ting Zhou, Bingyu Zhang, Dazheng Zhang, Qiong Wu, Jiajie Chen, Lu Li, Yiwen Lu, Michael J Becich, Saul Blecker, Nymisha Chilukuri, et al. Body mass index and postacute sequelae of sars-cov-2 infection in children and young adults. *JAMA network open*, 7(10):e2441970–e2441970, 2024.