



Causal Inference Methods for High-Resolution Data: Methodological Innovations for Estimating Treatment Effects in Complex Data Structures

Citation

Meng, Xiang. 2025. Causal Inference Methods for High-Resolution Data: Methodological Innovations for Estimating Treatment Effects in Complex Data Structures. Doctoral Dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://dash.harvard.edu/handle/1/42719741>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

HARVARD
Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Statistics

have examined a dissertation entitled

Causal Inference Methods for High-Resolution Data:
Methodological Innovations for Estimating Treatment Effects
in Complex Data Structures

presented by Xiang Meng

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature

A handwritten signature in black ink, appearing to read "N. Pillai", written over a horizontal line.

Typed name: Prof. Natesh Pillai

Signature

A handwritten signature in black ink, appearing to read "I. Bojinov", written over a horizontal line.

Typed name: Prof. Iavor Bojinov

Signature

A handwritten signature in black ink, appearing to read "L. Miratrix", written over a horizontal line.

Typed name: Prof. Luke Miratrix

Date: May 12, 2025

Causal Inference Methods for High-Resolution Data: Methodological Innovations for Estimating Treatment Effects in Complex Data Structures

A DISSERTATION PRESENTED
BY
XIANG MENG
TO
THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2025

© 2025 - *Xiang Meng*

ALL RIGHTS RESERVED.

Causal Inference Methods for High-Resolution Data: Methodological Innovations for Estimating Treatment Effects in Complex Data Structures

ABSTRACT

This dissertation develops statistical methods for high-resolution data across three distinct but complementary domains. In this context, "high-resolution data" refers to information structures with granular detail that traditional statistical methods often simplify or ignore—whether in the form of temporally dense decision points, precise treatment timing that defines the causal question itself, or detailed within-cluster distributions that standard approaches typically collapse into simple averages.

The first chapter evaluates a sequential risk time sampling algorithm implemented in a mobile health intervention study. This algorithm addresses the challenge of delivering interventions across 144 potential decision points per day while maintaining both treatment frequency and uniform distribution constraints. The analysis demonstrates that the algorithm successfully balances these objectives, enabling valid causal inference for identifying contexts where interventions prove beneficial in chronic disease management.

The second chapter integrates staggered adoption designs with survival analysis to estimate causal effects when treatment timing varies across subjects. The treatment variable in causal inference is problem-defining, and when it has granular temporal structure (like

exact transplant dates), it fundamentally reshapes the causal question rather than merely adding predictive power. By combining hazard-based modeling with double machine learning, this approach maintains robustness to model misspecification while providing interpretable treatment effect estimates. Applications to heart transplant data demonstrate superior performance compared to traditional methods, with extensions to business contexts where timing of customer actions has causal implications.

The third chapter introduces a variance estimator for matching methods that remains valid even when matched samples substantially overlap—a common challenge when treatment groups are small. While the granular within-cluster information has always existed in matched datasets, this approach uniquely leverages the full distribution of control outcomes within each matched set, outperforming existing methods in simulation studies and providing a generalized theoretical framework applicable to various matching procedures. The methodological advance enables more reliable inference for policy evaluations across economics, education, and public health.

Together, these three studies advance causal inference methodology by designing statistical approaches that properly leverage high-resolution data structures, providing researchers with practical tools for deriving valid insights from increasingly complex and detailed data while improving both empirical accuracy and decision-making relevance.

Table of Contents

TITLE PAGE	i
COPYRIGHT	ii
ABSTRACT	iii
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
ACKNOWLEDGMENTS	xiii
1 INTRODUCTION	1
2 CHAPTER 1: EVALUATION OF THE HEARTSTEPS ONLINE SAMPLING ALGORITHM	5
2.1 Introduction	5
2.2 The HeartSteps V2V3 Trial	12
2.3 The Sequential Risk Time Sampling (SeqRTS) Algorithm	15
2.4 Evaluation of Algorithm’s Performance on the Core Design Constraints . .	20

2.5	Identify key covariates to improve the online algorithm's performance on meeting the core design constraints	29
2.6	Conclusions and Discussion	43
3	CHAPTER 2: TIME-VARYING CAUSAL SURVIVAL LEARNING	49
3.1	Introduction	49
3.2	Problem Set Up	54
3.3	Review of Handling Time-Varying Treatment in Survival Models	61
3.4	Double Machine Learning Estimator	63
3.5	Simulation Study	67
3.6	Data Analysis – Stanford Heart Transplant dataset	73
3.7	Conclusion	78
3.8	Introduction	80
3.9	Problem Setup	83
3.10	The Inference Problem	88
3.11	The Standard Error Estimator	95
3.12	Simulation	105
3.13	Application of the Variance Estimator to Other Estimators	110
3.14	Conclusion	113
A	SUPPLEMENTARY MATERIALS FOR CHAPTER 1	115
A.1	The Sequential Risk Time Sampling Algorithm Parameters Used in the HS V2V3 Studies	115
A.2	A Coding Issue that Affects Evaluation	119
A.3	Checking Plausibility of the MCAR Assumption	121
A.4	Checking the assumption that the mean is constant across the days	123
A.5	An illustrative example of why block averages do not sum to the daily average	125
A.6	Discussion on Not Using KL Divergence for Uniformity	126

A.7	Details on managing baseline variables	128
A.8	Ensure No Endogeneity from the Time-Varying Covariates	130
A.9	The Robustness of the Algorithm’s Performance	133
B	SUPPLEMENTARY MATERIALS FOR CHAPTER 2	136
B.1	Derivation of the TV-CSL Estimator	136
B.2	Key Lemma for Deriving the Score Function	138
B.3	Proof of Proposition 1	140
B.4	Parametrizing the PDF for the Piecewise Cox Model	144
C	SUPPLEMENTARY MATERIALS FOR CHAPTER 3	148
C.1	Proof of Theorem 3.10.1	148
C.2	Proof of Lemma 3.11.1	152
C.3	Proof of Lemma 3.11.2	162
C.4	Other useful Lemmas	165
C.5	Compare the Lipschitz Condition to that in the Existing Litarature	168
C.6	Comparison with Theorem 1 of [125]	169
C.7	Full Simulation Settings	171
	REFERENCES	178

List of Figures

2.2.1	Decision times during a typical participant's day in the HeartSteps V2V3 study. $B_t = 1$ if the participant is sedentary at decision time t ; $I_t = 1$ if the participant's decision time t is available.	14
2.2.2	This figure consists of four histograms illustrating different aspects of participant activity in the study. From the top left to the bottom right: The first two histograms depict the distribution of the number of days each participant stayed in the study for V2 and V3, respectively. The third histogram showcases the distribution of the number of available and sedentary decision times per participant in V2, while the fourth histogram presents the distribution of the number of available and sedentary decision times per participant in V3.	16
2.4.1	Histogram of the average number of treatments per participant for each block and the whole day for all 82 participants. Vertical lines indicate the block and daily goals (0.5 for each block and 1.5 for the whole day). The error bars represent the point estimates and 95% confidence intervals of the average number of treatments across all participants.	25

2.4.2	Histogram of participant average uniformity (mean absolute difference, MAD) for each block and the entire day for all 82 participants. The error bars represent the point estimates and 95% confidence intervals of the MAD across all participants.	26
2.5.1	Decision tree for fitting df.drop (tree 1, on the left) and df.KNN (tree 2, on the right). Each node contains a number representing the average treatment for participants in the node, and the proportion of participants in the node. Both trees utilize sedentary behavior and activity level to make splits. Tree 1 was preferred due to its stability.	35
2.5.2	Plot of $\hat{g}_t$, the estimated remaining risk times, over t in a block. We plot $\hat{g}_t$ versus decision time in a block, varying two sets of values for R_t , the full length of time up until t that a participant is at risk, or the risk “run length”, one of two variables that $\hat{g}_t$ uses. The two sets of values are the lower and upper bounds of R_t : a) $R_t = 1$ for each decision time (the left plot), i.e., assuming the participant has not been sedentary up until t , and b) $R_t = t$, i.e., assuming the participant has been sedentary all the time prior to time t (the right plot). Each 12 decision times is an hour. These two plots show that $\hat{g}_t$ changes the most after 12 decision times, i.e., when t goes to a new hour. Specifically, $\hat{g}_t$ is smaller when t falls in the later hours of the block.	37
2.5.3	Predicted average treatment and MAD with respect to the most relevant features. Red horizontal lines are at 0.5 for average treatment and 0.01 for MAD, representing benchmarks to differentiate blocks with high and low average treatment and MAD values. Grey vertical lines represent plausible cutoffs to differentiate blocks. The three distinct lines in each plot represent the first quartile (Q1), median (Q2), and third quartile (Q3) of EndProp and HourVar values, respectively.	39

3.5.1	Comparison of Mean Squared Error (MSE) between singly robust (S-Lasso) and doubly robust (TV-CSL) methods under varying conditions. Each panel shows a distinct combination of baseline hazard (η_0) and propensity score specifications, with rows representing η_0 complexity (linear/complex) and columns indicating propensity score specification (correctly specified/mis-specified). Results are shown across sample sizes (200, 500, 1000, 2000) on the x-axis, with MSE on the y-axis. Red and blue lines represent S-Lasso and TV-CSL performance, respectively. Error bars indicate ± 1.96 Monte Carlo standard errors.	71
3.5.2	Comparison of Mean Squared Error (MSE) between S-Lasso and TV-CSL methods across different model specifications and sample sizes. Each panel represents a different combination of baseline hazard (η_0) and heterogeneous treatment effect (HTE) specifications, with rows indicating η_0 complexity (linear/complex) and columns indicating HTE complexity (linear/complex). The x-axis shows sample sizes (500, 1000, 2000), and the y-axis displays MSE. Red lines represent S-Lasso performance, while blue lines represent TV-CSL performance. Error bars indicate ± 1.96 Monte Carlo standard errors.	72
3.6.1	Distribution of Time of Heart Transplant	75
3.12.1	Empirical distribution of $\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{V^{-1/2}}$ versus the standard normal. . . .	107
3.12.2	Comparison of the Wild Bootstrap, A-E Bootstrap, and true values for confidence interval (CI) coverage and average CI length. The Wild Bootstrap method achieves shorter CIs but at the cost of lower coverage, while the A-E Bootstrap method closely approximates the true coverage but results in longer intervals. Horizontal dashed and dotted lines indicate the true coverage and CI length, respectively, for reference.	109

A.1.1	Fraction of time sedentary within and after current hour	117
A.1.2	Run lengths of being classified as either “Not Sedentary” or “Sedentary” respectively.	118
A.4.1	Plot of average of number of treatments over all study days for all blocks. The averages are taken over users who are available on the day.	124
A.4.2	Plot of average MAD over all study days for all blocks. The averages are taken over users who are available on the day.	125
A.5.1	This example illustrates why block goals (average of 0.5 treatments per block) are achieved, but the daily goal (average of 1.5 treatments per day) is not. Blue boxes represent blocks or days with at least one available time, and the numbers inside them show the total number of treatments in each block or day. On day 3 and 4, there is no available time in block 1, so the average for block 1 does not include these two days. However, these two days are included in the daily average calculation because there are available times in blocks 2 and 3.	126
A.9.1	Boxplot of the average number of treatments per participant for each block calculated using SeqRTS and the Oracle. SeqRTS_Sim represents the performance of SeqRTS on the simulated risk times, while SeqRTS_Actual represents the performance of SeqRTS on actual decision times in HSV2V3.	135
B.4.1	Comparison of treatment and control PDFs	145

THIS WORK IS DEDICATED TO ALL OF THE PEOPLE WHO LOVE AND CARE FOR ME.
THANK YOU.

Acknowledgments

I am deeply grateful to the many individuals and institutions who have supported me throughout my graduate studies and the completion of this dissertation. This journey would not have been possible without their guidance, encouragement, and generosity.

First and foremost, I extend my deepest gratitude to my thesis advisor, Natesh Pillai, for his unwavering support, guidance, and mentorship throughout my doctoral journey. I thank him for deciding to be my advisor when I needed one with little doubt, for helping me connect to collaborators, for advising my academic progress, and for checking into my mental health.

I am profoundly grateful to my committee members. To Luke Miratrix, thank you for helping me work on projects when I was seeking an advisor. I benefit greatly by observing how you ask research questions, and I appreciate the way you organize your lab by treating people as friends. Your support was a first step for me to get saved from academic probation. To Iavor Bojinov, thank you for your guidance. It's been an interesting journey as we step forward together. Your supportiveness has been key to my confidence, and I've learned a lot from observing your way of communication and running the group.

I owe special thanks to Susan Murphy, my first mentor and collaborator on the first

chapter. She taught me the invaluable lesson of making baby steps—a painful but necessary learning as I had been superficial and unrealistic before. I know she always has good intentions. I am also deeply thankful to Jun Liu, who made himself available to listen to my disappointments during academic probation and offered many useful insights. His support significantly reduced my anxiety.

For those I worked with in teaching, I extend my sincere appreciation. Kelly McConville helped me form my philosophy of pushing students to take risks comfortably rather than scaring them. James Xenakis gave me credit that I did not feel valued before. I also thank Giovanni Parmigiani, Joe Blitzstein, Xiao-Li Meng, and Alex Young for their guidance in my teaching journey.

I am deeply indebted to Jess Jackson for patiently helping me with paper edits countless times and teaching me how to write clearly. She also listened to me during the beginning of my hard change from not trusting the American culture to finding points to immerse myself in it. The same gratitude extends to the FWC folks (Chris, Josephine, Sam) for similar support.

To my research collaborators who taught me how to conduct research—Jonathan Che, Hilary Weingarten, Jonathan Huang, Thomas Richardson, Linbo Wang, Walter Dempsey, and Peng Liao—thank you for your patience and knowledge sharing.

I am grateful to the Bok Center team—Pamela Pollock, Adam Beaver, and Tamara Brener—for providing a space for me to grow as a teacher and exposing me to different disciplines. Thank you for listening to my ideas patiently, making me feel accepted.

The Statistics Department administrative staff—Emily Palmer, Lauryn Aliwalas, Dayanara Torres, and Kathleen Cloutier—deserve special thanks for their timely support and informal listening sessions that helped release my stress.

I would also like to acknowledge my mental health counselors—Janet, Carmen, and Mojgan—who played a very important role in my personal growth.

To my academic coach, Khaleem Ali, I owe tremendous gratitude for teaching me about

American culture and academic culture. I did not know how to play the game in academia and graduate school, and he essentially showed me how.

I extend my heartfelt thanks to my friends and colleagues in the Harvard Statistics Department, especially those in my cohort: James Bailie, Tyler Tianle Liu, Yanke Song, Biyonka Liang, Kuanhao Jiang, Yi Zhang, Jane Qizhao Chen, Yufan Li, Yicong Jiang, and Buyu Lin. Other department friends and lab friends include Anna Trella, Shuangning Li, Kyra Gan, Peng Liao, Yongyi Guo, Hsin-Yu Lai, and Raaz Dwivedi. Your camaraderie, collaboration, and support have made this journey not only productive but also enjoyable.

To my friend Yu-Sheng Wang—thank you for standing by me through my darkest moments. Your unwavering support and kindness carried me when I needed it most, and I will forever be grateful. To my Cambridge friends, Hang Jiang, Zhijian Ren, Chengfeng Mao—thank you for a decade of laughter, deep conversations, and steadfast encouragement. You reminded me of joy during stressful times and inspired me to keep going.

I would also like to acknowledge the funding sources that made this research possible. The second chapter of this dissertation was supported by grants from the NIH/NIDA P50 DA054039, NIH/NIBIB and OD P41EB028242, NIH/NHLBI R01HL125440, and NIH R01 GM152549. I am grateful for the financial support that allowed me to focus on my research and academic pursuits.

Finally, I thank my family for their unconditional love, encouragement, and sacrifices. Your belief in me has been a constant source of strength and motivation.

To all those who have supported me in ways big and small, I am profoundly grateful. This dissertation is as much a product of your kindness and guidance as it is of my own efforts.

1

Introduction

The rapid advancement of data collection technologies has transformed how researchers analyze and derive insights across diverse fields. From mobile health applications that capture moment-by-moment user behavior to e-commerce data that document detailed customer histories, we now have access to what can be termed “high-resolution data”—information characterized by granular temporal detail or multidimensional structure. While traditional statistical methods like t-tests can still be applied to such data, they often do so by reducing rich information structures to simplified binary treatments or aggregate outcomes, ignoring the detailed structure that makes the data valuable. These traditional methods were designed for low-resolution data and fail to utilize the full potential of

granular covariates, temporal patterns, or complex interdependencies present in modern datasets.

It is thus desirable to develop new methods that can properly utilize high-resolution data to eliminate bias that emerges from ignoring the granular structure, improve efficiency by accounting for the data’s full complexity, or both. This dissertation presents a collection of three studies that advance the methodological rigor and practical utility of causal inference by responding to the challenges posed by such granular data, with applications spanning business decision-making, healthcare delivery, and social science research. Though examining different applications, these chapters are united by their focus on developing robust statistical approaches that represent important steps toward fundamental improvements in domain practices.

Chapter 1 evaluates a sequential risk time sampling (SeqRTS) algorithm implemented in the HeartSteps mobile health intervention study. While not a causal inference methodology per se, this chapter examines an experimental design algorithm critical for enabling valid causal inference in subsequent analyses. The high-resolution aspect here lies in the decision times—up to 144 potential intervention points per day for each participant—where the algorithm must determine whether to deliver an anti-sedentary message. The chapter demonstrates that SeqRTS successfully satisfies two key constraints: maintaining an average of 1.5 interventions per day and ensuring uniform distribution of interventions across decision times. This uniformity is essential for proper context sampling, allowing researchers to estimate causal effects across various participant states without introducing bias from uneven intervention distribution. Beyond merely improving existing methods, this represents a fundamental shift toward personalized intervention strategies in chronic disease management—where traditional approaches were limited to infrequent clinical visits, high-resolution data now enables the discovery of daily contexts where targeted nudges can effectively promote lasting behavioral change, with the ultimate goal of finding specific circumstances where interventions prove most beneficial.

Chapter 2 bridges staggered adoption designs from econometrics with survival analysis to address a fundamental challenge in medical research: estimating causal effects when treatment timing varies across patients. The high-resolution component appears in the continuous treatment timing variable—the exact date when a patient receives an intervention like heart transplantation. Traditional causal inference methods typically assume fixed treatment assignment, failing to capture crucial timing differences—a patient who waited 10 days for a donor heart likely differs substantially from one who waited 300 days in terms of disease progression, organ compatibility, and overall health status. Using only a binary treatment indicator ignores this temporal variation and can lead to significant bias in effect estimation. This chapter develops a framework that integrates hazard-based modeling with double machine learning to estimate heterogeneous treatment effects while maintaining robustness to model misspecification. Through both simulations and application to heart transplant data, the approach demonstrates superior performance compared to traditional methods, enhancing hazard estimation in transplant studies by properly accounting for time-varying treatments. This improvement allows medical professionals to more accurately estimate and communicate transplant risks to patients and their families—a critical advancement in medical communication and clinical decision-making. The methodology extends naturally to business contexts—for example, estimating whether purchasing a low-cost digital item increases the likelihood of a subsequent high-value purchase, with direct implications for designing consumer incentives.

Chapter 3 addresses a critical gap in statistical inference for matching estimators, focusing on scenarios with substantial overlap between matched samples. When the same control units appear repeatedly across multiple matched sets, complex dependencies arise that existing inference procedures fail to properly account for. This chapter introduces a variance estimator that maintains validity even under high overlap conditions, outperforming state-of-the-art methods in simulation studies. Additionally, it develops a generalized theoretical framework that relaxes standard Lipschitz continuity assumptions, allowing for

broader applicability in settings where the derivative grows moderately but not uniformly. This advancement is particularly valuable for evaluating initiatives with limited adoption, such as education policy interventions where small treatment samples often limit statistical inference. By developing more robust variance estimation procedures for matching methods, this work improves the reliability of policy impact assessments when traditional methods like bootstrapping fail, providing essential tools for uncertainty quantification in high-stakes decision-making across economics, epidemiology, and policy evaluation.

Together, these three chapters provide a cohesive set of methodological innovations addressing key challenges in high-resolution causal inference. While each chapter focuses on a distinct problem within a specific domain, the underlying statistical principles and methodological approaches contribute to a unified framework for handling complex data structures in causal inference. These studies share a common methodological vision: they aim to sharpen the accuracy and interpretability of causal estimates by designing methods that align closely with the data structure and scientific objectives of the problem at hand. This alignment is critical not only for improving empirical insights but also for making causal methods more trustworthy and actionable in real-world decision-making contexts. By developing these methods across business, healthcare, and social science applications, this thesis offers researchers practical tools to derive valid causal insights from the increasingly detailed and complex data that characterize modern empirical research.

Through this work, I hope to contribute to a growing body of research that takes seriously both the opportunities and perils of modern data—advancing causal inference tools that are adaptable, reliable, and guided by practical relevance.

2

Chapter 1: Evaluation of the HeartSteps Online Sampling Algorithm

2.1 INTRODUCTION

Micro-randomized trials [MRTs; 34, 66, 77] are experimental studies that sequentially randomize each participant at thousands of decision times over the course of a given trial. MRTs are frequently used in digital intervention development [18, 19, 37, 46, 49, 69, 70, 79, 84, 88, 89, 99, 100, 109, 115, 122]. Digital interventions involve the use of smart devices such as a smartphone, smartwatch, or other wearable device to deliver sequences

of treatments to individuals as they go about their daily lives. These treatments comprise messages with varying content, such as motivational messages, nudges of different types, reminders, and so on.

Multiple MRTs have been conducted in the process of developing the HeartSteps intervention package designed to encourage physical activity (Klasnja et al. 2019, Liao et al. 2020, Spruijt-Metz et al. 2022). One intervention component in the package, called the anti-sedentary intervention, is the delivery of messages that are designed to interrupt sedentary behavior (to stand up, stretch, and/or move around). In the HeartSteps V2V3 MRTs [HSV2V3; 79, 109], decision times occur at five-minute intervals over the course of each study day. An individual who has taken fewer than 150 steps in the 40 minutes prior to a decision time is considered sedentary. Anti-sedentary messages were randomized to be sent or not at sedentary decision times throughout the day. For a decision time t , we use B_t to denote sedentariness: $B_t = 1$ indicates that time t is a sedentary time, and $B_t = 0$ indicates otherwise.

The anti-sedentary intervention in the HSV2V3 MRTs needs to meet two common MRT design constraints: (1) the *average treatment constraint*, which restricts the average number of treatments delivered to participants in a pre-specified time range, e.g. a four-hour time block in a study day. Such constraints can help reduce the burden to participants of digital interventions (Klasnja et al. 2008, Figueroa et al. 2021, Liao et al. 2018); and (2) the *uniformity constraint* aims to ensure that treatments are distributed uniformly across times allowed for randomization. This is to ensure context sampling rates align with the frequency at which the contexts occur during sedentary times. A context is information concerning an individual in which the treatment might have an effect (e.g., whether an individual is stressed or not) [87]. Since the distribution of the context is not known beforehand, treatments must be uniformly distributed across times allowed for randomization to ensure no context is favored. Otherwise, it is possible that the number of treatments delivered in certain context (e.g., when an individual is stressed) is low

(when stressed sedentary times occur fairly frequently). As a result, the ability to learn the causal effect in this context is harmed since we have little post-treatment data for that block [35, 78].

Both design constraints can be satisfied by a straightforward MRT design (i.e., randomization probabilities) when the total number of sedentary times is known in a pre-specified time range. HSV2V3 consists of 12-hour days segmented into three equal four-hour blocks, each comprising 48 decision times. The scientific team aims to ensure that on average for each participant, 0.5 interventions are delivered uniformly per four-hour block during sedentary times ($B_t = 1$) with none happening during non-sedentary times ($B_t = 0$). Suppose beforehand that 20 of 48 decision times are sedentary times. In this case, randomizing the “send a message” option at each sedentary time with a probability of $\frac{0.5}{20} = 0.025$ will ensure both constraints are met: The average treatment constraint is met because the treatment probabilities sum up to 0.5, the desired average. The uniformity constraint is met because the treatment probabilities are equal to each other. However, the number of sedentary times in each block for each participant is completely unknown at the beginning of the block, so this strategy cannot be applied.

To address this challenge, an online algorithm called Sequential Risk Time Sampling (SeqRTS) was developed [35, 78]. In simulated data, SeqRTS was shown to meet the two constraints well on average across participants. However, the question of whether the algorithm can successfully meet the two constraints when implemented in the real world is still uncertain, as success on a simulated dataset does not guarantee success in the real world due to the existence of uncontrolled factors [83, 114]. A randomized trial is the gold standard for determining an algorithm’s efficacy in mobile health and, in general, the entire healthcare sector when making real-time interventions [74]. SeqRTS was implemented in the HSV2V3 study to manage the randomization of anti-sedentary messages with the aim of meeting the two constraints. We conducted an evaluation using data collected from HSV2V3 to assess the algorithm’s performance in meeting these

constraints.

Our analysis shows that the algorithm successfully met both constraints on average across all participants in HSV2V3. We also show that it did not meet the ideal goal of achieving both constraints for each individual participant. This is unsurprising to the algorithm’s designers because this is one of the first few studies in which this algorithm has been implemented in a real study. We identified participants and blocks where the algorithm underperformed and found baseline and time-varying variables that characterize them. After identifying these participants and blocks, we provide suggestions to move the algorithm’s performance closer to the ideal goal of meeting both constraints for each individual participant.

SeqRTS is designed to perform well even with moderate effectiveness of the anti-sedentary messages. Therefore, if the messages result in mildly reduced sedentary behavior, the algorithm should still function effectively, and there is no need to adjust the evaluation to account for days with low sedentary behavior.

Our contributions are summarized as follows:

1. **Evaluation of a Real-World Implementation of an Online Algorithm** We present a comprehensive study of 82 participants over 9,600 days, demonstrating that the SeqRTS algorithm performed well in the real-world setting of an MRT for meeting the design constraints. We select appropriate metrics and discuss conditions under which the evaluation is fair. This is the first extensive evaluation of such an algorithm in this context.
2. **Methodological Contributions:** While the number of interventions delivered is an obvious metric to assess SeqRTS’s performance on achieving the average treatment constraint, the appropriate measure for evaluating the algorithm’s performance on the uniformity constraint is less obvious. We discuss possible metrics for uniformity and propose using the Mean Absolute Deviation (MAD) as a suitable measure.

The MAD has advantage over KL divergence, the uniformity metric suggested in previous work by [78].

3. **Suggestions for Future Improvement of the Algorithm:** We identify key variables that can improve the algorithm’s performance closer to the ideal goal of meeting both constraints for each individual participant. Specifically, we identify variables correlated with SeqRTS’s under-performance and quantify their relationship to the algorithm’s under-performance. We then discuss improvement strategies utilizing these findings, thereby offering actionable insights for further refinement of the algorithm.

This study demonstrates the effectiveness of SeqRTS in meeting the two design constraints in a real-world setting and provides a methodological blueprint for future algorithmic advancements in similar domains.

The organization of the paper is as follows. In Section 2.2, we introduce the HeartSteps V2V3 trial. In Section 2.3, we describe the SeqRTS algorithm used in this trial. In Section 2.4, we evaluate the algorithm’s performance on the average treatment constraint and the uniformity constraint. In Section 2.5, we investigate baseline variables and time-varying covariates that could be used to improve the algorithm’s performance, and propose improvement strategies for the algorithm. We conclude the paper by discussing limitations and future work in Section 2.6.

2.1.1 RELATED WORK

The first category of related work pertains to the literature where sedentary times are of interest. Sedentary times are instances of broadly defined risk times, which are times of elevated stress or heightened risk for participants. Assessing treatment effects during risk times is of interest to many investigators. An example besides HSV2V3 is Sense2Stop [17], a smoking cessation trial. In Sense2Stop, risk times are times when a participant (a

smoker) is stressed. An investigator is interested in assessing the effect of a message to reduce the probability of smoking at risk times. Risk times are infrequent and occur in a stochastic manner. To handle the inherently random temporal distribution of risk times, [35] defined the stratified micro-randomized trial (sMRT), where strata are times when a participant is at different levels of risk. In the anti-sedentary intervention of HSV2V3, risk times are sedentary times, and the number of strata is two (sedentary times and non-sedentary times). In general, sMRTs may encompass more than two strata.

The second category of related work focuses on existing evaluations of online algorithms. Online algorithms are designed for scenarios in which data or inputs are received incrementally over time [133]. In these circumstances, decisions are made in real-time as data point arrives, usually with incomplete knowledge or a limited understanding of potential future events. The decision problem that we have introduced here is an online problem because the arrival of future sedentary times is unknown.

Many works on evaluations of online algorithms exist. There are theoretical analyses that focus on comparing the performance of an online algorithm to that of an ideal offline counterpart with complete foresight [competitive analysis; 15, 80], or aim to establish theoretical bounds on the regret, the difference between the algorithm’s performance and that of an optimal decision-making policy [12, 14, 16, 73, 128]. Our work differs from the above because it is an empirical evaluation.

Our work also differs from previous empirical evaluations. Empirical studies of online algorithms often rely on simulations to generate outcomes, even when using real-world datasets [7, 121]. Our evaluation differs because we evaluate the algorithm on the trial where it was implemented, Evaluating algorithms on real data is crucial for understanding their practical effectiveness and robustness in real-world scenarios [24]. In real-world applications, online algorithms have been employed in studies like those by [85] and [120], which utilized online multi-armed bandit algorithms to make optimal allocation of service calls from health workers to prevent drops in participant’s engagement in a healthcare

study. They aimed to ensure that the best possible decisions were made with regard to an objective function. Our work has a different focus; we concentrate on ensuring that constraints are met.

The third category of related work focuses on the notion of uniformity in experimental design. In terms of managing the uniformity, the closest comparison in the literature is the biased coin design (BCD), first proposed by [Efron \(1971\)](#) and later extended by [Wei \(1978\)](#). The BCD appears in a two-arm (one treated, one control) randomized control trial (RCT) setting. In BCD, the notion of uniformity is the balance between the number of treated subjects and the number of controls. The biased coin design (BCD) aims to maintain a good balance by assigning a lower probability to the treated arm when the number of treatments exceeds the number of controls at the treatment assignment time, and vice versa.

The similarity between the balance in BCD and uniformity in SeqRTS is that they both tackle a decision problem where future event knowledge is limited (i.e., an online decision problem): in BCD, the future number of participants is unknown, and in SeqRTS, the future number of sedentary times is unknown. The key difference lies in the quantity that they aim to maintain well. The quantity that the BCD aims to maintain well is the ratio of treated participants to control participants. Keeping a good balance requires that this ratio is as close to one as possible. For SeqRTS, the quantity to maintain well is the variance of the treatment probabilities across the sedentary times (see Section 2.4.2 for a detailed discussion). The uniformity constraint requires the variance to be as low as possible, ideally zero. Unlike BCD, it is not a concern if the number of treated times does not equal the number of control times.

2.2 THE HEARTSTEPS V2V3 TRIAL

HeartSteps is a mobile intervention developed to encourage physical activity. HeartSteps has multiple intervention components (referred to as the “package” in the Introduction). In this paper we focus on the anti-sedentary intervention component that aims to disrupt the sedentary behavior by providing suggestions to stand up, stretch, and/or move around via the participant’s smartphone. We use a “participant” or a “user” interchangeably to refer a person enrolled in the study. HeartSteps V2 and V3 (HSV2V3) involved participants who were newly diagnosed with stage 1 hypertension (a diastolic pressure of 80 to 89 mm Hg or a systolic pressure of 130 to 139 mm Hg, [Mayo Clinic 2021](#)). The primary difference between HeartSteps V2 and HeartSteps V3 was the study duration. In V2, the duration of each participant’s participation was 90 days, running from May 2019 to January 2020, whereas in V3, it extended to 270 days, beginning in October 2019 and ending in February 2021. Note the participants entered the study at various dates so the whole span of the study is longer than 90 days for V2 and 270 days for V3.

In HeartSteps V2V3, all participants were provided a Fitbit tracker and the HeartSteps mobile phone application. Each intervention is delivered as a notification on the participant’s smartphone via the HeartSteps phone app. We refer to these notifications as the treatment, and in the case of the anti-sedentary intervention, the treatment was an anti-sedentary message. Each participant was asked to wear the Fitbit tracker and to install HeartSteps software on their tracker, which we call the HeartSteps watch face. The watch face was used to transmit step count data at short intervals in time to the HeartSteps cloud server.

A baseline period is a period during which data is collected on the participant’s activity without any intervention. The HeartSteps interventions start the day after the baseline period ends. The baseline period ends after 7 days of baseline information is collected. Sometimes it takes more than 7 days for some participants to finish the baseline period

because participants may not wear the watch on some days. The median length of the baseline period is 11 days.

On each study day, the eligibility for anti-sedentary messages is calculated for a twelve-hour period whose end time is one hour after the participant’s usual post-dinner time (collected before the beginning of the day). This eligibility is determined every five minutes, referred to as decision times. At each decision time, a participant was deemed either “unavailable for treatment” (see below for a definition) or “available for treatment”. If available for treatment, the participant was classified in one of two strata, currently sedentary or not currently sedentary, on the basis of the step count data over the last 40 minutes from the HeartSteps cloud server. Participants were considered sedentary if they had taken fewer than 150 steps in the prior 40 minutes. HeartSteps only randomized the delivery of the anti-sedentary messages at times when participants were both available and sedentary. We call these decision times “risk times.” Both availability and sedentariness at decision time t are determined right before time t .

A participant can be unavailable for various reasons. First, participants are considered unavailable if they have been active recently, that is, if they have taken more than 2,000 steps in the previous 120 minutes. Second, participants are considered unavailable if the watch face is not connected to the cloud server. This includes both decision times at which the participant is not wearing the Fitbit wrist tracker and decision times when the participant is wearing the Fitbit but the watch face is unable to communicate with the cloud based server. If the cloud based server does not have any records from the watch face for the 40 minutes prior to a decision time, the decision time is defined as unavailable. Third, decision times are considered unavailable if the participant has been sent any type of notification (including an anti-sedentary message) in the prior hour. Finally, participants were able to adjust settings on the HeartStep app to indicate that they do not want to be disturbed. All decision times occurring during these do not disturb times are considered unavailable.

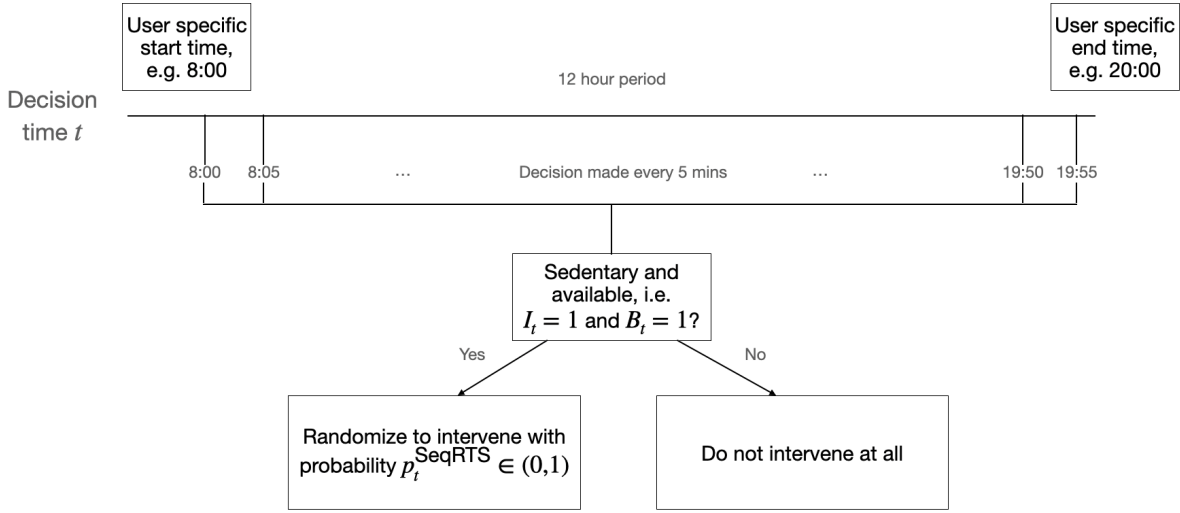


Figure 2.2.1: Decision times during a typical participant's day in the HeartSteps V2V3 study. $B_t = 1$ if the participant is sedentary at decision time t ; $I_t = 1$ if the participant's decision time t is available.

Unfortunately, when storing the data, only one binary availability indicator was kept for the aforementioned four criteria, instead of four separate binary variables indicating availability for each criterion. This data storage issue creates challenges when analyzing the algorithm's performance, primarily because the third criterion causes availability to depend on the algorithm implemented in the trial. We discuss the implications of this issue in Section 6.

The two HeartStep studies recruited 113 adults; 19 registered but did not participate with the result that 94 participants participated in the combined HSV2V3 study. Data from a further 12 participants were excluded due to technological issues. In particular, 9 of 94 participants (9.6%) are excluded due to the watch face not getting set up correctly, as it was not possible to determine whether a decision time was a risk time (with participant both available and sedentary). 3 of 94 participants (3.2%) are excluded due to no Fitbit data. In summary, we use the data from 38 participants in V2 and 44 participants in V3.

See Figure 2.2.1 for a typical participant's day. The 82 participants were in the study (meaning from the first day they became eligible for randomization to the last day before

they leave the study) for a total duration of 11,316 participant days, leading to 1,629,504 ($11,316 \text{ days} \times 144 \text{ decision times per day}$) decision times from June 2019 to February 2021. Unfortunately, due to server-side errors from 15 February 2020 to 9 April 2020, step count data from the watch face was not transmitted to the cloud server; thus, 272,592 decision times (or 1,893 participant days) are not included. Before 15 February 2020, 82 participants contributed for 5,523 participant days. From April 9 to end of study, 36 participants contributed for 3,900 participant days. Together these two periods resulted in 1,356,912 decision times (or 9,423 days). Of the 1,356,912 included decision times, participants were unavailable for treatment or not sedentary during 1,190,170 decision times (87.7% of the time). Hence, data associated with 166,742 risk times (available and sedentary decision times) will be analyzed here. Across the 166,742 risk times, 10,137 anti-sedentary messages were delivered.

Figure 2.2.2 presents histograms of the number of days each participant stayed in the study for both V2 and V3, as well as histograms of the number of available and sedentary decision times per participant for V2 and V3. We notice that the modal number of days in the top row, which illustrates the histograms of the number of days per participant, does not align with the expected 90 days for V2 and 270 days for V3 participants. This discrepancy arises for two main reasons. Firstly, participants might leave the study early. Secondly, due to study consent constraints, data is retained for a maximum of 90 days for V2 and 270 days for V3 participants. This includes the baseline period when the intervention did not start. Another observation is that in the bottom row, the distribution of available and sedentary decision times is uneven across participants.

2.3 THE SEQUENTIAL RISK TIME SAMPLING (SEQRTS) ALGORITHM

The “Sequential Risk Time Sampling” Algorithm (SeqRTS) (Liao et al. 2018, Dempsey et al. 2020) was designed to satisfy two constraints, the average treatment constraint and

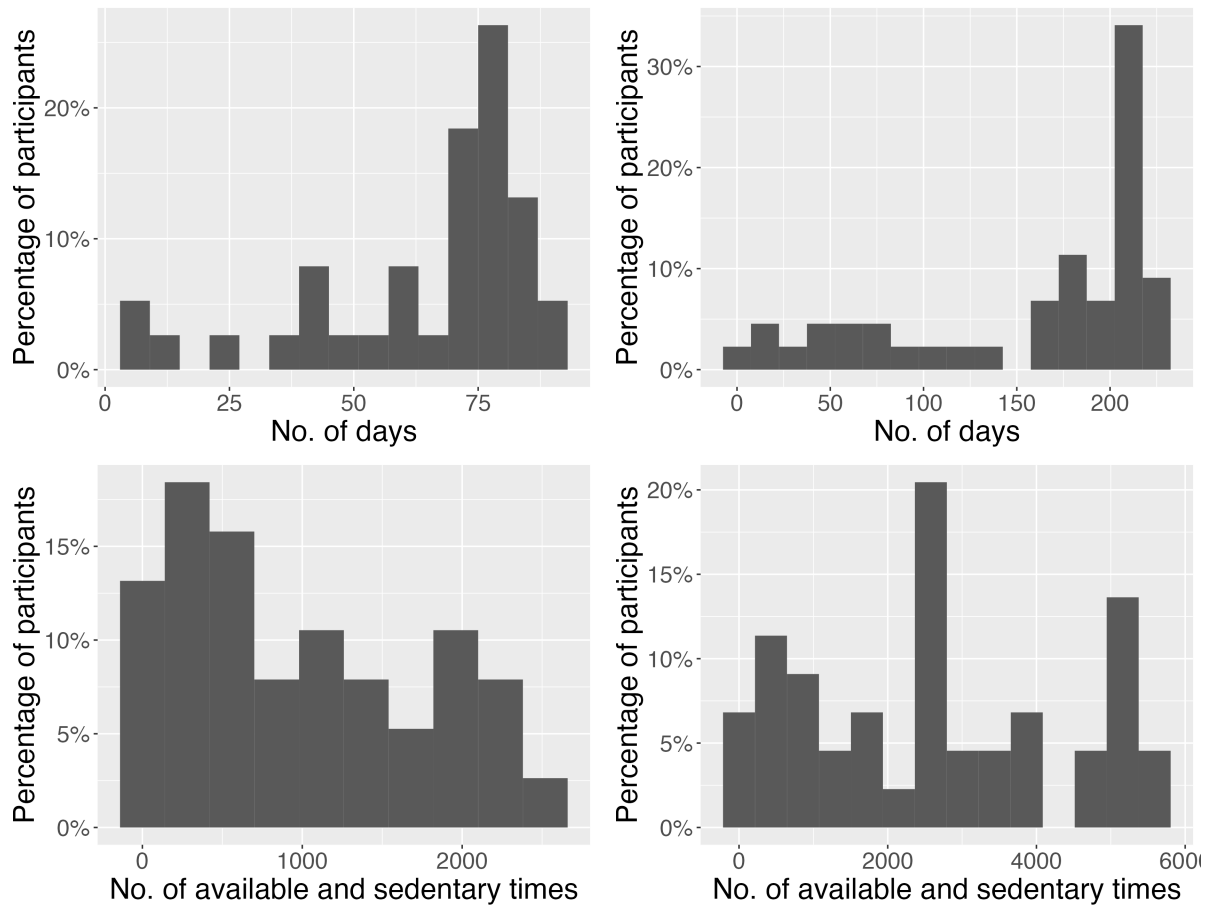


Figure 2.2.2: This figure consists of four histograms illustrating different aspects of participant activity in the study. From the top left to the bottom right: The first two histograms depict the distribution of the number of days each participant stayed in the study for V2 and V3, respectively. The third histogram showcases the distribution of the number of available and sedentary decision times per participant in V2, while the fourth histogram presents the distribution of the number of available and sedentary decision times per participant in V3.

the uniformity constraint. In HeartSteps V2V3, the average treatment constraint was an average of 1.5 anti-sedentary messages per day. This number was chosen by the scientific team to reduce burden on participant [41, 65, 78]. Each day is segmented into three equal four-hour blocks. SeqRTS works at the block level, and the goal of an average of 1.5 anti-sedentary messages per day translates into an average of 0.5 anti-sedentary messages per block. The uniformity constraint means the treatments should be delivered uniformly across all risk times. In this paper, we measure uniformity by checking how closely the randomization probabilities within each block resemble each other. See Section 2.4.2 for our specific choice of the Mean Absolute Deviation.

We first introduce some notations before introducing SeqRTS formally. For $j = 0, 1, 2, 3$, let $\mathcal{T}_{i,j,d}$ denote the set of decision times for participant i , in block j ($j = 0$ means for the whole day), and on day d . In HeartSteps V2V3, $\mathcal{T}_{i,0,d} = \{1, 2, \dots, 144\}$, $\mathcal{T}_{i,1,d} = \{1, 2, \dots, 48\}$, $\mathcal{T}_{i,2,d} = \{49, 50, \dots, 96\}$, $\mathcal{T}_{i,3,d} = \{97, 98, \dots, 144\}$. Thus, $|\mathcal{T}_{i,0,d}| = 144$ and $|\mathcal{T}_{i,j,d}| = 48$ for $j = 1, 2, 3$. We then use $T_0 = 144$ and $T = 48$ to denote number of decision times in a day or a block. Even though the beginning times in block 2 and 3 are not 1, we use $t = 1, \dots, T$ to count the times in all three blocks. Denote $B_{i,j,d,t} \in \{0, 1\}$ and $I_{i,j,d,t} \in \{0, 1\}$ as indicators of sedentary behavior and availability at time t , for participant i , in block j , and on day d , where $B_{i,j,d,t} = 1$ represents the participant is sedentary, $I_{i,j,d,t} = 1$ represents the participant is available. A risk time is when participant is both sedentary and available, that is, $B_{i,j,d,t} = 1$ and $I_{i,j,d,t} = 1$. We denote the indicator of the risk time as $X_{i,j,d,t}$, and denote the total number of risk times in each block by $N_{i,j,d} = \sum_{t=1}^T X_{i,j,d,t}$. We denote time-varying covariates as $Z_{i,j,d,t}$. To simplify notations, we will often suppress the i, j, d subscripts and refer to the indicators primarily as B_t , I_t , X_t , Z_t unless the context necessitates the specification of the participant, block, or day. We also use bar notation to denote the vector of variables. For example, $\bar{Z}_t = \{Z_1, Z_2, \dots, Z_t\}$.

We now use $N := \sum_{t=1}^T X_t$ as the simplified notation for $N_{i,j,d}$, the number of risk times. When N is known for a block at all risk times $\{t : X_t = 1\}$, the following sampling

probability will achieve both constraints:

$$p_t^{Oracle} = \frac{0.5}{N} \quad (2.1)$$

This is because this probability is equal across all risk times, so the uniformity constraint is satisfied, and the probabilities sum up to 0.5, so the average treatment constraint is satisfied. We call Equation 2.1 the oracle probability. Note that although the oracle probability will achieve both constraints, it is impractical because the number of risk times, N , is unknown at risk time t . Specifically, $N = \sum_{t'=1}^T X_{t'} = \sum_{t'=1}^{t-1} X_{t'} + \sum_{t'=t}^T X_{t'}$. The first term, $\sum_{t'=1}^{t-1} X_{t'}$, is the number of risk times occurred before t . It is known at time t . The second term, $\sum_{t'=t}^T X_{t'}$, is the number of remaining risk times after time t , which we denote as g_t . g_t is unknown at time t , which makes N unknown.

If g_t is known, an alternative strategy that would achieve both constraints is as follows: We first refer to the average treatment allowed in the block as the “budget,” which we represent by N_0 ($N_0 = 0.5$ in HSV2V3). At each decision time t , we denote the remaining “budget” by $N_0 - \sum_{s=1}^{t-1} X_s \cdot p_s$, where p_s is the treatment probability at time s . At time t , if we divide the remaining “budget” by the remaining risk times g_t , the probability will be equal to the oracle probability (Equation 2.1). This is the strategy that SeqRTS takes. See [78] for a more detailed walkthrough.

Practically, g_t is not known at time t , so it needs to be estimated $\hat{g}_t$. We represent the estimator of g_t as $\hat{g}(H_t)$, where H_t denotes the set of time-varying variables collected so far in the block, i.e., $H_t = \{\bar{Z}_t, \bar{A}_{t-1}, \bar{p}_{t-1}, \bar{I}_t, \bar{B}_t, \bar{X}_t\}$. To manage the estimation errors associated with $\hat{g}_t$, we make other quantities flexible. Although $N_0 = 0.5$ represents the concrete goal of the average number of treatments to be sent in a block, we make it a tuning parameter, $\hat{N}_0$, to cope with the uncertainty of $\hat{g}$, ensuring that SeqRTS functions effectively. At decision time t in the j -th block on day d , SeqRTS determines the probability of delivering treatment as follows:

$$p_t^{SeqRTS} = \frac{\hat{N}_0 - \sum_{s=1}^{t-1} X_s \cdot p_s^{SeqRTS}}{1 + \hat{g}(H_t)}. \quad (2.2)$$

In the parameter estimation stage, $\hat{g}$ was first estimated, then $\hat{N}_0$ was tuned given $\hat{g}$. The estimation and tuning for $\hat{g}$ and $\hat{N}_0$ were done using data from the prior study, Heartsteps V1 [HSV1, 67]. $\hat{g}$ and $\hat{N}_0$ were never updated during HSV2V3.

The time-varying inputs used by $\hat{g}$ are a subset of H_t . These inputs include: (1) **Hour of the decision time:** $\text{Hour}_t := \lfloor t/12 \rfloor$; (2) **Risk “run length”:** R_t represents the full length of time up until t that a participant is at risk. R_t can be defined recursively through the formula below:

$$R_t := \begin{cases} 0 & \text{if } X_t = 0, \\ R_{t-1} + 1 & \text{if } X_t = 1, \end{cases}$$

where R_1 is initialized to 0 to accommodate the beginning of the block. In Appendix A.1, we provide the exact form of $\hat{g}(H_t)$ used in HSV2V3.

$\hat{N}_0$ was tuned by selecting a value of $\hat{N}_0$ that satisfies two design constraints as closely as possible given a fixed $\hat{g}$. The best $(\hat{N}_0, \hat{g})$ pair ¹ was selected, resulting in $\hat{N}_0 = 1.8$. One reason to set a $\hat{N}_0$ that is much greater than 0.5 is that $\hat{g}$ was generally too high, resulting in a higher denominator in Equation 2.2 ². Consequently, those blocks would generate fewer treatments than the target number of 0.5 if $\hat{N}_0$ were set to a lower value. Lastly the probability, p_t^{SeqRTS} is clipped to belong to the interval $[0.005, 0.2]$ to ensure the output value remains within a reasonable range and to avoid extreme probabilities that might lead to under- or over- treatment. The algorithm’s developer chose 0.2 as an

¹the pair under which simulated data based on HSV1 achieved an average treatment of 0.5 per block and minimized KL divergence between the treatment probabilities and the oracle probabilities. See Section 2.4.2 for a formal definition.

²See Appendix A.1.3 for details

upper bound based on an analysis of risk times in HSV1 data. Appendix A.1 provides details for the choice of the choice.

Since $\hat{N}_0$ and $\hat{g}_t$ were overestimated, we are uncertain about the algorithm’s performance in practice. [78] suggests it should perform well on average across participants. In the next section, we will evaluate whether this expectation holds true.

2.4 EVALUATION OF ALGORITHM’S PERFORMANCE ON THE CORE DESIGN CONSTRAINTS

In this section, we evaluate of SeqRTS’s performance on the two core design constraints in HSV2V3, the average treatment constraint and the uniformity constraint. We first discuss the choice of data to use in evaluating the algorithm’s performance. We then define measures of average treatment and uniformity. Finally, we evaluate the SeqRTS algorithm’s ability to achieve an average of 1.5 treatments per day and its performance on maintaining uniformity across participants.

2.4.1 SELECTION OF BLOCKS AND DAYS

Before conducting our evaluation, we select blocks and days for each participant to accurately represent the algorithm’s performance. Below, we mainly describe the procedure for selecting entire days, but the same criteria apply to the selection of blocks.

There are decision times, and indeed, entire days when participants are unavailable (see Section 2.2 for criteria for a participant’s unavailability). Denote the availability indicator of a study day d as $I_d = (\sum_{t=1}^{T_0} I_{d,t} \geq 1)$, where there is at least one available decision time (with some abuse of notation, as we use I to represent both the availability of a decision time t and the availability of a study day d). For a participant, days we choose for evaluation are days that the participant is available: $\mathcal{D} = \{d : I_d = 1\}$. This selection of days results in 6,327 participant days out of the total 9,423 eligible participant days.

Days when a participant is not available are those in which $I_d = 0$. They are excluded from our analysis because there is no data for the algorithm’s performance on these days (since it could not perform). This means I_d an indicator of missing data. Inference based on the observed data alone might be biased if data is missing not at random [MNAR; 82]. Here we argue that our selection of days does not suffer from this issue.

Recall the unavailability reasons described in Section 2.2. When the participant is not available on a day, it means the participant was physically active, did not wear the watch, or that the watch face was not connected to the cloud server. We are not interested in days when a participant is physically active (e.g., when they go on a hike) because it makes no scientific sense to provide any anti-sedentary message when someone is active. We are also not concerned with days when a participant did not wear the watch (e.g., due to busy schedules, travel, or simply forgetting), as there is no opportunity to assess sedentary behavior on those days. It is important to note that days when a participant does not wear the watch may be influenced by an unobserved variable U (e.g., mental tiredness). For example, if a participant is mentally tired, they may choose not to wear the watch and may also exhibit more sedentary behavior. On such days, the algorithm might not perform well due to $\hat{g}_t$ being an inaccurate prediction.

As a result, the algorithm could appear to perform well by omitting days when a participant is mentally tired and not wearing the watch. We acknowledge this limitation, but note that our analysis of the HSV2V3 study is representative of Kaiser Permanente patients newly diagnosed with stage 1 hypertension. The algorithm’s performance is evaluated marginally across variables like U , and therefore reflects the general behavioral patterns of the study population, including variability in watch-wearing behavior.

The days we are interested in are those when participants wear the watch and are not active all day. Missingness of the data on these days means the watch is not connected to the cloud server, which is caused by technical malfunctions of wearable devices, such as syncing issues, software glitches, i.e., technical reasons instead of user’s choices. Specif-

ically, these reasons are not related to unmeasured factors like a participant’s mental state or unmeasured sedentary status on that day. Hence, we argue this missingness is at random (MAR), as it is either unrelated to the study variables or can be explained by observed data.

We apply the same logic to our choice of blocks. The number of days for which we have data from blocks 1, 2, and 3 are 5,575, 5,473, and 4,994 respectively. Note that the number of included days for analysis of each block is smaller than the total of 6,327 participant days selected for analysis; if a participant day has an available decision time in block 1 but none in blocks 2 and 3, this day is included in the evaluation of block 1 and that of the day, but not in the evaluation of blocks 2 and 3.³ We denote $\mathcal{D}_j$ as the set of blocks j in which a participant has an available decision time for $j = 1, 2, 3$. When $j = 0$ (whole day), we denote $\mathcal{D}_0 = \mathcal{D}$

During the data analysis, we discovered a coding issue: for some decision times (most of them in block 3), a failure in correctly calculating Hour_t , the hour of the decision time t (due to a coding issue in timezone conversion), caused $\hat{g}(H_t)$ to be calculated incorrectly. As a result, at these decision times, the algorithm only produced a randomization probability of 0.005, the lower bound, which is very different from the correctly calculated randomization probabilities. Including too many of these decision times would improperly represent the algorithm’s performance on meeting the two constraints. We exclude 5.4% of the blocks where there were many decision times at which $\hat{g}_t$ was incorrectly calculated. The detailed exclusion criterion is presented in Table 4 in Appendix A.2, where this issue is described in more detail.

³See Section A.5 for a detailed illustration

2.4.2 MEASURES OF NUMBER OF TREATMENTS AND UNIFORMITY

On day d , at each decision time t , the SeqRTS algorithm calculates a randomisation probability $p_{d,t}$ as detailed in Equation 2.2. An action $A_{d,t}$ is sampled from $p_{d,t}$ through $A_{d,t} \sim \text{Bern}(p_{d,t})$. The total number of treatments on day d is calculated as $Y_d = \sum_{t=1}^{T_0} A_{d,t}$. This Y_d is used to evaluate the algorithm’s performance on achieving 1.5 anti-sedentary messages per day. Note that we are implicitly summing over risk times, since $p_{d,t} > 0$ only at risk times.

In this paper, “uniformity” refers to how evenly treatments are distributed across all risk times. A sampling algorithm that achieves uniformity should generate treatment probabilities that are as similar as possible. Therefore, we measure uniformity using treatment probabilities. An ideal uniformity metric is zero when all probabilities are equal and increases as differences among probabilities grow. One such metric is the mean absolute difference (MAD), which measures the average absolute difference between probabilities at risk times and the mean probabilities on a day d :

$$V_d = \frac{\sum_{t: X_{d,t}=1} |p_{d,t} - \frac{1}{N_d} \sum_{t': X_{d,t'}=1} p_{d,t'}|}{N_d} \quad (2.3)$$

A smaller value of MAD represents more uniform probabilities. A perfect uniformity, i.e., when probabilities are equal on day d , means $V_d = 0$.

In past work, [78] proposed evaluating uniformity using Kullback-Leibler (KL) divergence between SeqRTS probabilities and the oracle probabilities (Equation 2.1) at each risk time. The KL divergence on a day d is an average of the KL divergence at all risk times in the day:

$$V_d^{KL} = \frac{\sum_{t: X_{d,t}=1} KL(p_{d,t}, p_{d,t}^{Oracle})}{N_d} \quad (2.4)$$

where $KL(p, q) = p \log \left(\frac{p}{q} \right) + (1-p) \log \left(\frac{1-p}{1-q} \right)$, and $p_t^{Oracle} = 0.5/N_d$. They use $V_d^{KL} = 0$ to represent uniformity. However, $V_d^{KL} \neq 0$ even when uniformity is achieved. A toy

example illustrates this phenomenon: Consider a block with two risk times. The oracle probabilities are $0.5/2 = 0.25$ for both risk times. Assume the SeqRTS probabilities are both 0.1 at the two risk times. The uniformity is achieved in this block since these two probabilities are equal, but they fail to meet the average treatment constraint, as their sum is 0.2, differing from the block’s average treatment constraint of 0.5. A calculation shows that the MAD, V_d equals zero, but the KL divergence $V_d^{KL} = 0.031$, is greater than zero. This demonstrates that KL divergence is affected by whether the average treatment constraint is met, in addition to the uniformity constraint. We discuss this observation in more detail in Appendix A.6.

2.4.3 EVALUATE THE ALGORITHM’S PERFORMANCE ON THE TWO CONSTRAINTS OVER ALL PARTICIPANTS

OBJECTIVES, ASSUMPTIONS AND THE METHOD

The algorithm was designed to achieve, ideally, on average, 1.5 treatments per day with uniformity for each participant. This is a difficult goal to meet because the SeqRTS algorithm implemented in HSV2V3 applied a single set of $(\hat{g}, \hat{N}_0)$ to all participants and was never updated during the study, making it likely that the algorithm performs differently across participants. This is supported by the data.

In Figure 2.4.1, we plot histograms of $\bar{Y}_{i,j} = \frac{1}{|\mathcal{D}_j|} \sum_{d \in \mathcal{D}_j} Y_{i,j,d}$, the average number of treatments sent on the available days and blocks $j = 0, 1, 2, 3$ over all participants $i = 1, 2, \dots, 82$. We also plot histograms of $\bar{V}_{i,j} = \frac{1}{|\mathcal{D}_j|} \sum_{d \in \mathcal{D}_j} V_{i,j,d}$, the average MAD for all three blocks and the whole day, in Figure 2.4.2. These histograms show that the algorithm’s performance varies among participants, indicating that the two constraints were not met for each participant.

A useful goal would be to assess the algorithm’s performance in achieving these two constraints when averaged over all participants. This goal is important because if the

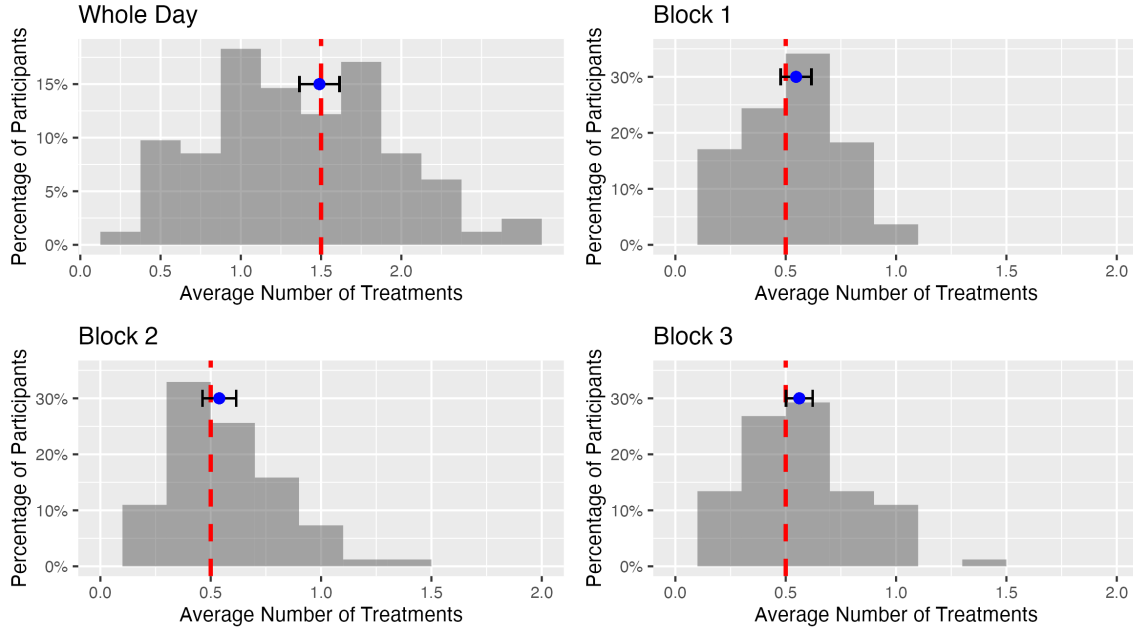


Figure 2.4.1: Histogram of the average number of treatments per participant for each block and the whole day for all 82 participants. Vertical lines indicate the block and daily goals (0.5 for each block and 1.5 for the whole day). The error bars represent the point estimates and 95% confidence intervals of the average number of treatments across all participants.

algorithm achieves it, it demonstrates significant potential. In the rest of this section, we will investigate this using a statistical model.

To assess the goal, we perform inference on average number of treatments per day and average MAD on the whole population. These averages are over days when participants wear the watch and are not active all day (see Section 2.4.1 for a discussion). Note that the number of treatments and the MAD for each participant i , $Y_{i,d}$ and $V_{i,d}$, are outcomes that vary over days d . We fit generalized estimating equations [GEE; 36, 76] to account for the longitudinal nature of the data.

We assume the population mean is constant throughout the study. An exploratory analysis shows that the number of treatments and MAD, when averaged over participants, do not vary across days. A test of significance on the study day variable using the same regression model as introduced later in Equation 2.5, but with the study day term

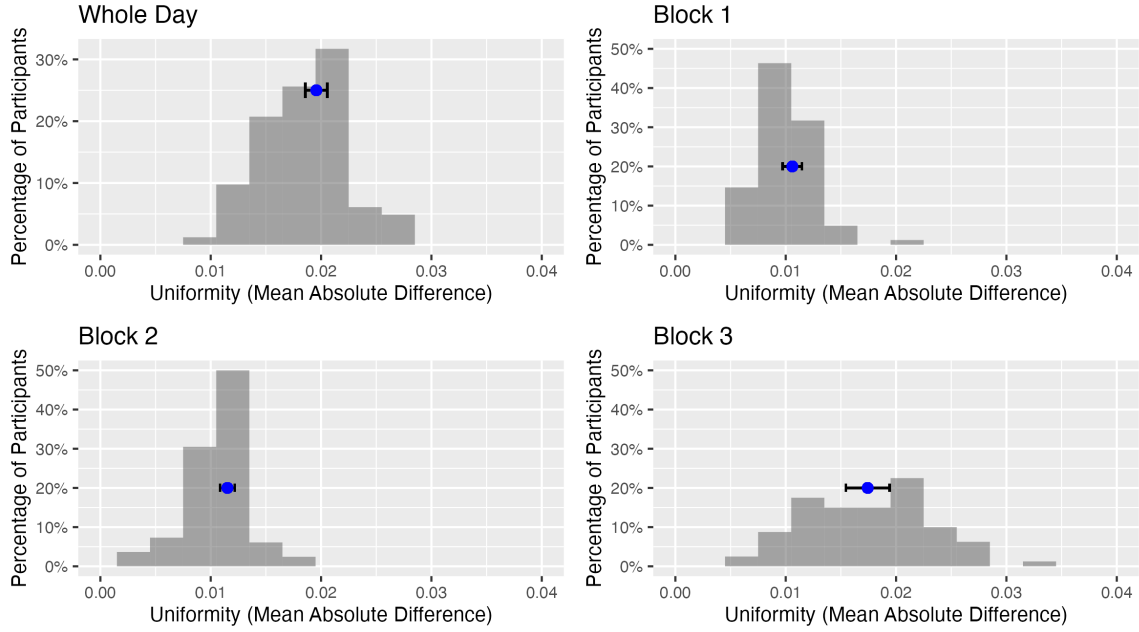


Figure 2.4.2: Histogram of participant average uniformity (mean absolute difference, MAD) for each block and the entire day for all 82 participants. The error bars represent the point estimates and 95% confidence intervals of the MAD across all participants.

added, further supports this assumption. We provide detailed analyses in Appendix A.4. Additionally, we assume the treatment on the previous day does not affect sedentariness on the present day. This is supported by a data analysis in Appendix A.8.2, where we showed that there is insufficient evidence to indicate that anti-sedentary messages sent on one day have a causal effect on sedentariness the next day, when controlling for past sedentary behavior.

We now describe the inference procedure for the average treatment. A similar procedure is applied on the inference for the uniformity. We use the log mean model as our mean model to account for the count nature of $Y_{i,d}$. The model is presented in Equation 2.5. Here, μ_0 represents the marginal mean, which is the average number of treatments sent across all participants. η is an over-dispersion parameter to account for the possibility that the data do not exactly follow a Poisson distribution (note that log mean model implies a Poisson regression). For simplicity, when presenting the results, we report only

the standard error $\sqrt{\hat{\eta}\hat{E}[Y_{i,d}]}$ and not $\hat{\eta}$ (where the hat notation represents estimators).

$$\begin{aligned}\log E[Y_{i,d}] &= \log \mu_0 \\ \text{Var}(Y_{i,d}) &= \eta E[Y_{i,d}]\end{aligned}\tag{2.5}$$

We fit the model using both an independent working correlation model and a first-order autoregressive model (AR-1) model. The AR-1 model accounts for the correlation between outcomes on different days, where outcomes closer in time are more correlated than those further apart. It is particularly useful for longitudinal data, where outcomes (number of treatments and MAD) may be correlated within each participant and across days. The AR-1 assumption does not need to be correct, but the estimation will be more efficient when it is correct.

For MAD $V_{i,d}$, we fit the GEE using an identity link to estimate the mean MAD across all participants. This is because uniformity is a continuous variable, and exploratory analysis shows that the histogram of MAD for each participant across all days is approximately bell-shaped.

Recall from Section 2.3 that the algorithm operates at the block level. Besides the daily goals, we also analyze whether the algorithm achieves the block goals. The analyses for each block and the whole day were conducted separately. We conclude the section by noting that the validity of the GEE results relies on the assumption that data is missing completely at random (MCAR). Appendix A.3 shows that this is a plausible assumption.

RESULTS

We now present the evaluation results for average treatment and uniformity. The first two columns in Table 2.4.1 present the estimation results on average number of treatments across users. We make two observations. First, the 95% confidence interval covers the

goals (0.5 or 1.5), except for block 3, where the average is slightly greater than the goal of 0.5⁴. Second, both AR-1 and independence models produce similar estimation results. The AR-1 correlation coefficient is significant (p-values for all blocks, which are omitted from the table, are less than 0.05), indicating that outcomes in consecutive days are likely correlated. Given this plausible correlation between consecutive days, we present the results using the AR-1 model in the following sections of the paper.

It is a significant result for SeqRTS to achieve the average treatment constraint across all users because the third unavailability criterion—unavailability for one hour after each treatment—created additional difficulty. The difficulty is created because the time when a treatment is delivered is an unknown event, even to the oracle, who knows the number of risk times. When a treatment is delivered, the number of risk times could be reduced, and hence the total number of treatments delivered in a block could be lower than 0.5 for the oracle. Indeed, in Appendix A.9, we use a simulation to show that on all available blocks, the oracle algorithm delivers an average number of treatments lower than the goal of 0.5 across all participants. This demonstrates that SeqRTS outperforms the oracle approach in handling real-world complexities.

The last two columns in Table 2.4.1 present the estimation results on average MAD across users. Blocks 1 and 2 have an average MAD of approximately 0.01, but block 3 exhibits a higher MAD of approximately 0.018. The larger MAD in block 3 is due to the presence of risk times with a lower bound probability of 0.005, which differs significantly from other probabilities. This is a result of the coding issue mentioned in Section 2.4.1 and detailed in Appendix A.2. As described in Section 2.4.1, days with large proportions

⁴Note that the block averages do not sum up to the daily average. This discrepancy arises from the possibility that a participant is available during one block but not the others on a specific day. Consequently, the daily average, which accounts for a greater number of days, is lower than the sum of the block averages. A detailed illustrative example can be found in Appendix A.5.

Table 2.4.1: Average treatment and uniformity for each block and the entire day across all 82 participants, estimated by generalized estimating equations (GEE). Each cell contains the point estimate and the 95% confidence interval (presented in the bracket).

Block / Working Correlation	width=			
	Average Number of Treatments		Uniformity (MAD)	
	Independence	AR-1	Independence	AR-1
Whole Day	1.490 (1.365, 1.615)	1.486 (1.365, 1.607)	0.019 (0.018, 0.020)	0.019 (0.018, 0.020)
Block 1	0.547 (0.477, 0.617)	0.552 (0.483, 0.621)	0.010 (0.009, 0.011)	0.010 (0.009, 0.011)
Block 2	0.539 (0.462, 0.615)	0.544 (0.469, 0.618)	0.011 (0.010, 0.011)	0.011 (0.010, 0.011)
Block 3	0.562 (0.501, 0.622)	0.569 (0.511, 0.627)	0.017 (0.016, 0.019)	0.017 (0.016, 0.019)

of coding errors were omitted, but days with mildly large proportions of coding errors were retained, primarily affecting block 3 but not blocks 1 and 2.

Unlike the average treatment constraint, where a target is established, determining whether MAD is low is relatively subjective. Note that $MAD = 0.01$ for a block means that, on average, treatment probabilities within the block differ from their block average by only 0.01, which we interpret as good uniformity. $MAD = 0.018$ also indicates good uniformity, especially considering that the probabilities sum up to 0.5 on average. As such, average MAD across all participants of 0.01 (for blocks 1 and 2) and 0.018 (for block 3) indicates good uniformity.

2.5 IDENTIFY KEY COVARIATES TO IMPROVE THE ONLINE ALGORITHM'S PERFORMANCE ON MEETING THE CORE DESIGN CONSTRAINTS

The previous section highlighted the algorithm's good performance in meeting both the average treatment and uniformity constraints across all participants. However, the algorithm's performance did not meet the ultimate objective of fulfilling these constraints for each individual participant: as Figure 2.4.1 and 2.4.2 show, there are participants for whom the algorithm's performance is suboptimal.

In this section, we discuss which baseline (Section 2.5.1) or time-varying covariates

(Section 2.5.2) could be used to improve to algorithm’s performance in meeting constraints for each individual participant. We will do so by figuring out which covariate correlates with the algorithm’s performance metrics (average treatment and MAD) in HSV2V3. To figure out these variables, we first come up with candidate variables: for baseline variables, the scientific team suggested a list of candidates. For time-varying variables, we examine specific aspects of SeqRTS, namely the formula itself and factors that leads to variation of $\hat{g}_t$ that affects performance on achieving two constraints. After we have a list of candidate variables, statistical models are fit to both select baseline covariates from the candidate and identify the population for whom the current algorithm under-performed. For baseline variables, we use a Classification and Regression Trees [CART; 23]. For time-varying covariates, we use GEE [76] to identify the blocks for which the current algorithm under-performed (Section 2.5.2, 2.5.2). Last but not least, we make suggestions on directions for the algorithm’s improvement after identifying the covariates.

2.5.1 USING CART TO IDENTIFY THE POPULATION FOR WHOM THE CURRENT ALGORITHM UNDER-PERFORMED

DESCRIBE CANDIDATE BASELINE COVARIATES

The scientific team identified candidate variables thought to be correlated with sedentary behavior, and thus could be correlated with algorithm’s performance. These variables include 1. **gender**: captures the respondents’ self-identified gender based on the options provided in the questionnaire. The collected data for the “gender” variable contains responses of “female” or “male.” 2. **is retired**: Whether a participant is retired. 3. **sedentary behavior**: How many hours a participant typically spends sitting on a week-day in the past 7 days. 4. **activity level**: median daily total number of steps during the baseline period.

To ensure interpretability of the fitted model, we categorize numeric values with mean-

ingful cutoff. We categorize activity level into “Low Activity”, “Moderate Activity”, “High Activity”, and “Very High Activity” categories, as delineated by [117], using the cut-offs 5000, 7500, 9000. We also categorize sedentary behavior into “Low Sedentary”, “Moderate Sedentary”, “High Sedentary” using cutoffs of 4 and 8 sedentary hours in a day [20, 94]. Details on the rationale why and how we do this are discussed in Appendix A.7.1.

There are 9 participants for whom sedentary behavior data in the baseline covariates are missing. To address this missingness, we applied three methods—K-Nearest Neighbor (KNN) imputation, median imputation, and omission of missing entries—resulting in the `df_KNN`, `df_median`, and `df_drop` datasets, respectively. Details of how missingness was handled can be found in Appendix A.7.2.

USING CART TO SELECT BASELINE COVARIATES

The model we used to both select baseline covariates and identify the population for whom the current algorithm under-performed are the tree models: Classification and Regression Trees (CART) and random forests [22]. The tree models are selected in this analysis because they offer interpretable classification criteria. This is important for distinguishing participants for which the algorithm under-performed because each leaf node will be a sub-population with different performance.

The analysis proceeds in two steps. First, we conduct a stability check using leave-one-out cross validation (LOOCV) to select hyper-parameters, in other words, select a model to use. Then, with the selected model, we fit to the data and interpret the result. We conduct analysis on average treatment constraint and uniformity constraint separately. Since the analysis procedures are the same, we mainly present the analysis procedure for average treatment constraint and just report the result for the uniformity constraint.

First step: stability check We focus on the analysis of average treatment. For average treatment, we use $\bar{Y}_{i,j}$, average treatment for participant i in block j as the outcome,

for all $j = 0, 1, 2, 3$ To find an optimal model, we employ leave-one-out cross validation (LOOCV) to select the hyper-parameter. The rpart package in R [113] mitigates CART’s over-fitting by utilizing the complexity parameter (cp). We vary the fitting of trees with cp values ranging from 0.01 to 0.07. We also fit random forests, which have fewer over-fitting concerns, using lengths between 3 and 7. The resulting nodes symbolize participant clusters that share similar treatment averages.

In Table 2.5.1, we report the LOOCV training and testing mean squared error (MSE) and the average split count for all combinations fitted on the three datasets, df_KNN, df_median, and df_drop. The table presents results for average treatment, with a focus on interpreting the model’s stability. For the CART model, cp values of 0.05 and 0.07 demonstrate greater stability, showing a narrower gap between training MSE and validation MSE, which suggests less overfitting compared to cp values of 0.01 or 0.03. Additionally, the average number of nodes is lower for cp values of 0.05 and 0.07, indicating a simpler model structure. These conclusions hold across all three datasets.

For the random forest model, setting maxnodes to 3 achieves closer training and validation errors, comparable to the CART model with cp values of 0.05 and 0.07, especially for the df_drop and df_KNN datasets. However, the df_median dataset shows a slight discrepancy, where the CART model with cp values of 0.05 and 0.07 has a slightly higher validation error than the random forest with maxnodes set to 3, possibly due to sub-optimal imputation of missing data in the df_median dataset. In conclusion, we limited the number of nodes to around 3 to 4.

To ensure stable model performance, we selected a cp value of 0.07 for fitting the tree model, using only the df_drop and df_KNN datasets. These datasets exhibited a narrower gap between training MSE and validation MSE during the stability checks, making them ideal candidates for this phase. In contrast, the df_median dataset was excluded due to concerns about potential inaccuracies from sub-optimal imputation of missing data, as

Table 2.5.1: Prediction results for on Leave-One-Out Cross-Validation (LOOCV) for datasets created from three imputation methods (df_drop, df_KNN, df_median) over two interpretable models (CART and random forest) with various hyper-parameters. CART is evaluated with complexity parameters (cp) ranging from 0.01 to 0.07, and random forest with maxnodes ranging from 3 to 7. Panel A reports the training MSE, Panel B reports the validation MSE, and Panel C shows the mean number of nodes resulted in the models.

	CART (cp)				random forest (maxnodes)		
	0.01	0.03	0.05	0.07	3	5	7
<i>Panel A: Training MSE</i>							
df_drop	0.197	0.212	0.220	0.222	0.263	0.225	0.200
df_KNN	0.183	0.198	0.214	0.215	0.259	0.226	0.204
df_median	0.201	0.218	0.224	0.226	0.260	0.227	0.206
<i>Panel B: Validation MSE</i>							
df_drop	0.300	0.312	0.300	0.299	0.302	0.282	0.276
df_KNN	0.335	0.350	0.313	0.313	0.295	0.278	0.273
df_median	0.312	0.335	0.344	0.338	0.297	0.281	0.275
<i>Panel C: Mean number of nodes</i>							
df_drop	10.0	5.67	4.03	3.86	3.00	5.00	7.00
df_KNN	11.1	7.02	3.98	3.88	3.00	5.00	7.00
df_median	10.3	4.90	3.90	3.63	3.00	5.00	7.00

indicated by the larger gap between training and validation MSE.

Second step: Select meaningful sub-population where the algorithm’s performance is different After selecting the model, we fit the tree model with a cp of 0.07 on both the `df_drop` and `df_KNN` datasets and interpret the results. The decision tree is depicted in Figure 2.5.1. The outcomes of tree fitting on the `df_drop` and `df_KNN` datasets resulted in the formulation of two distinct trees: tree 1 and tree 2. To further assess stability, we randomly excluded 4 participants across 5 different random seeds and refit the data on both `df_drop` and `df_KNN`. The `df_drop` dataset consistently yielded tree 1, while the `df_KNN` dataset also produced tree 1 in 4 out of 5 instances, instead of tree 2. As a result, tree 1 was selected as the final model due to its demonstrated stability.

Tree 1 categorizes participants into three distinct groups, characterized by average treatment values of 1.1, 1.4, and 2. These groups can be intuitively labeled as under-treated, well-treated, and over-treated, respectively. The splitting variables are activity level and sedentary behavior. The first splitting criterion is based on activity level: if a participant has high or very high activity (daily step counts equal to or exceeding 7500), they are under-treated, receiving an average of 1.1 messages per day. This makes sense, as active participants are less available and thus have fewer opportunities to be randomized. The second criterion is based on sedentary behavior: if a participant does not have high or very high activity and is highly sedentary (sedentary hours equal to or exceeding 8 hours), they will be over-treated, receiving an average of 2 messages per day. This is logical, as being more sedentary provides more opportunities for randomization and, consequently, more treatments.

BASELINE VARIABLES ARE NOT CORRELATED WITH UNIFORMITY

We repeat the same procedure with the uniformity data. The result indicates that none of the baseline variables are predictive of uniformity, demonstrating that the uniformity

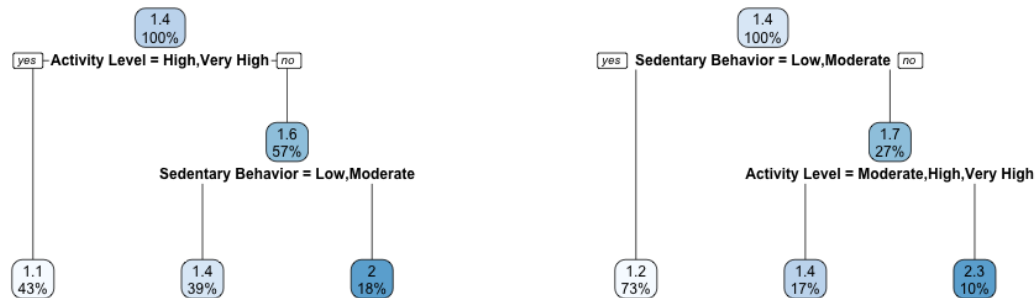


Figure 2.5.1: Decision tree for fitting `df_drop` (tree 1, on the left) and `df_KNN` (tree 2, on the right). Each node contains a number representing the average treatment for participants in the node, and the proportion of participants in the node. Both trees utilize sedentary behavior and activity level to make splits. Tree 1 was preferred due to its stability.

is not influenced by the baseline variables we considered.

2.5.2 USING GEE TO IDENTIFY INFLUENTIAL TIME-VARYING COVARIATES

In this section, we identify time-varying covariates that could be used to improve the algorithm’s performance in meeting constraints for each individual participant. We do so by figuring out which covariates correlate with average treatment and MAD on all 9,964 participant blocks. We use Generalized Estimating Equations (GEE) to conduct analysis. We present the analysis of the average treatment constraint and the uniformity constraint in Sections 2.5.2, 2.5.2 respectively.

IDENTIFY TIME-VARYING COVARIATES TO IMPROVE AVERAGE TREATMENT

In this section, we explore potential time-varying covariates that may improve SeqRTS’s performance in achieving an average treatment number of 0.5 per block for each participant. We first identify and construct time-varying covariates that are correlated with

average treatment over blocks, then fit a GEE regression model to examine their relationship with the number of treatments sent in the block. Lastly, we analyze the regression results and make a recommendation on how to improve SeqRTS. We construct these time-varying covariates by examining the numerator and denominator of the SeqRTS formula (Eq. 2.2) separately.

First, we examine the numerator. The numerator, $\hat{N}_0$, the approximate average treatment allowed in the block, is the same across all blocks. In Section 3, the explanation of how SeqRTS works shows that the sum of treatment probabilities in a block depends on how much of $\hat{N}_0$ is spent. For blocks with more risk times, SeqRTS has more opportunities to spend $\hat{N}_0$. Conversely, fewer risk times offer fewer opportunities to spend $\hat{N}_0$. This suggests that “riskness” is a relevant time-varying covariate to use in the regression. We refer to “riskness” as “sedentariness” in the remaining part of the section, since sedentariness is a more descriptive term for a participant’s behavior. Formally, we define mean sedentariness $\text{MeanSed} = \sum_{t=1}^T \tilde{X}_t$ (recall $T = 48$ represents the number of decision times in a block). We hypothesize that larger MeanSed is positively correlated with the average treatment.

Second, we examine the denominator $\hat{g}_t$, the estimated remaining risk times after time t . Our construction of $\hat{g}_t$, as illustrated in Figure 2.5.2, shows that $\hat{g}_t$ is smaller when t falls in the later hours of the block. The reason $\hat{g}_t$ is smaller near the end of the block is that there are fewer decision times near the end. When $\hat{g}_t$, the denominator is small, the treatment probability at time t is large. This suggests another time-varying covariate: end block sedentary proportion, or EndProp: Let N be the number of risk times in block. Then EndProp is the fraction of these risk times appearing at one of the last N decision times in a block. Mathematically, $\text{EndProp} = \frac{\sum_{t=T-N+1}^T \tilde{X}_t}{N}$. EndProp quantifies how much of the risk times appear at the end segment of a block. We hypothesize that the larger EndProp is, the larger the number of treatments is on average.

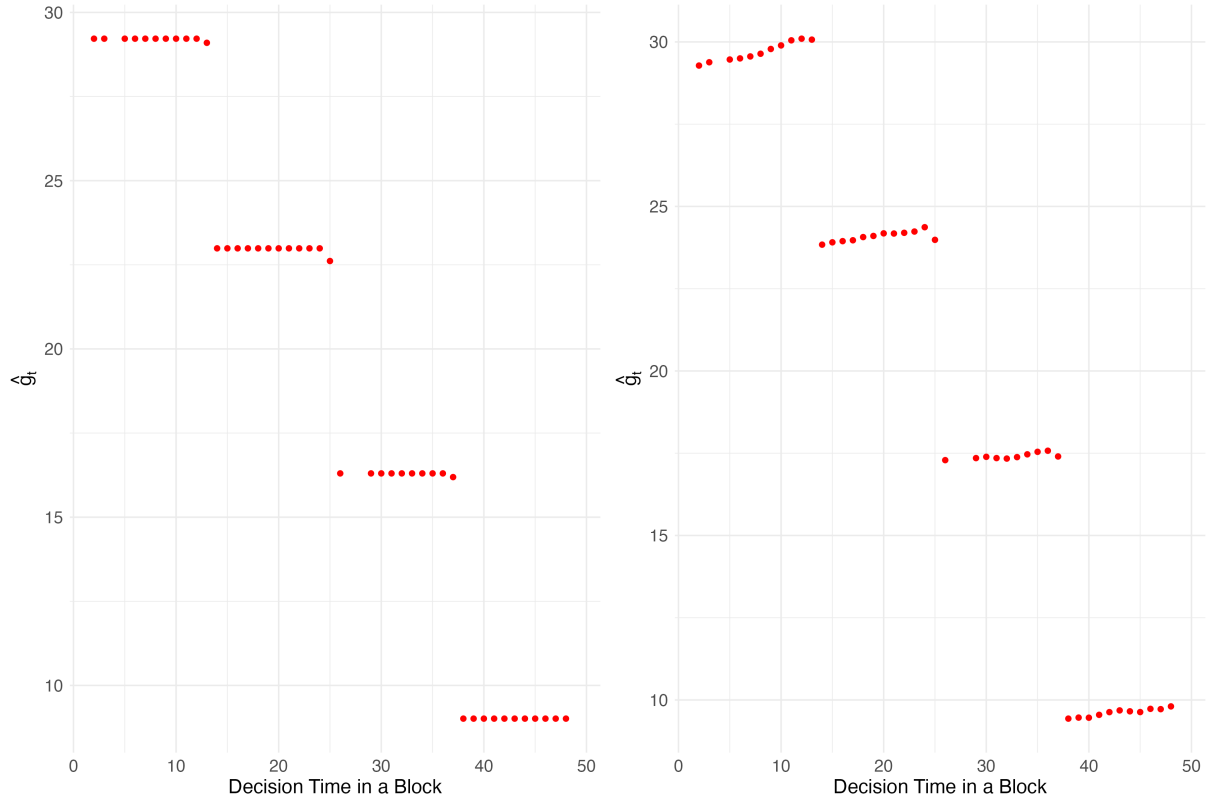


Figure 2.5.2: Plot of $\hat{g}_t$, the estimated remaining risk times, over t in a block. We plot $\hat{g}_t$ versus decision time in a block, varying two sets of values for R_t , the full length of time up until t that a participant is at risk, or the risk “run length”, one of two variables that $\hat{g}_t$ uses. The two sets of values are the lower and upper bounds of R_t : a) $R_t = 1$ for each decision time (the left plot), i.e., assuming the participant has not been sedentary up until t , and b) $R_t = t$, i.e., assuming the participant has been sedentary all the time prior to time t (the right plot). Each 12 decision times is an hour. These two plots show that $\hat{g}_t$ changes the most after 12 decision times, i.e., when t goes to a new hour. Specifically, $\hat{g}_t$ is smaller when t falls in the later hours of the block.

Before conducting the GEE regression, we took a few steps to ensure that there is no endogeneity, where endogeneity refers to the correlation between the predictors and the errors that can undermine the validity of the GEE analysis by giving biased and inconsistent estimates [36]. First, note that the MeanSed and EndProp are functions of the risk time indicator X_t . They are is susceptible to endogeneity: due to the one-hour unavailability following a treatment (the third cause of unavailability in Section 2.2), blocks with at least one treatment may have a different risk time distribution compared to blocks without any treatment, even if they have the same degree of sedentariness. As a result, we impute an availability indicator based on all other reasons for unavailability, excluding the third cause, by setting the user to be available for all decision times within one hour following a treatment ⁵. The corresponding risk indicator is denoted as $\tilde{X}_t$. Second, we use the prior 5-day average of the time-varying covariates instead of their same-day values. This is to remove reverse causality and ensure that the number of treatments does not influence the two time-varying covariates, thus ensuring no endogeneity in the regression. Lastly, we assume the treatment has no effect on sedentary behavior when controlling for past sedentary behavior. This assumption is supported by a data analysis in Appendix A.8.2.

We now use GEE regression to examine the relationship between MeanSed, EndProp and the number of treatments in a block, Y . An exploratory data analysis that considers all first and second-order terms of MeanSed and EndProp reveals that MeanSed, MeanSed², and EndProp are correlated with Y . We then fit a GEE with a log-linear mean model (to accommodate the count nature of the response variable Y) to examine the significance of MeanSed, MeanSed², and EndProp, and to quantify their relationships with Y . We implemented various correlation structures, including 'independent', 'AR-1', and 'exchangeable'. The results were consistent across these different specifications. We

⁵Details can be found in Appendix A.8.1.

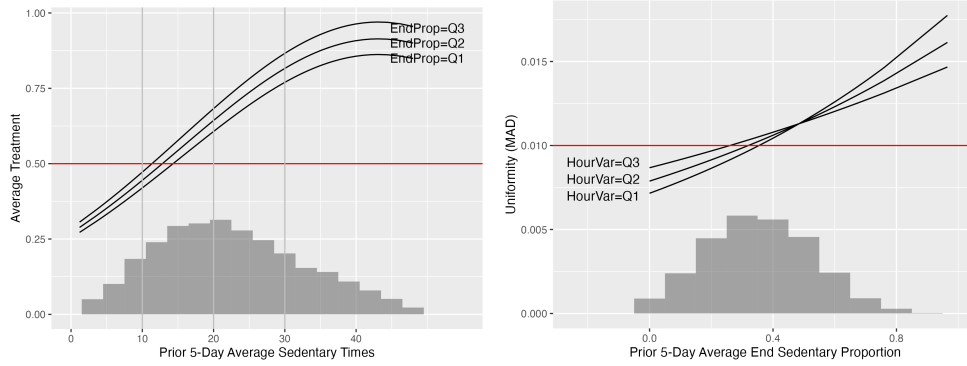


Figure 2.5.3: Predicted average treatment and MAD with respect to the most relevant features. Red horizontal lines are at 0.5 for average treatment and 0.01 for MAD, representing benchmarks to differentiate blocks with high and low average treatment and MAD values. Grey vertical lines represent plausible cutoffs to differentiate blocks. The three distinct lines in each plot represent the first quartile (Q1), median (Q2), and third quartile (Q3) of EndProp and HourVar values, respectively.

choose 'AR-1' because it is a plausible correlation structure for the outcome, as discussed earlier in Section 4.3. To check residuals, we set aside the last 20% of days as testing data for each participant, then fitted a locally estimated scatterplot smoothing (LOESS) curve between the fitted values and the Pearson residuals [36]. The residuals were closely aligned with zero on the residual plots for both the training and testing datasets. The residual plots of the GEE model shows no evidence of a relationship between the fitted values and the residuals.

Predictor	Estimate	Std. Err	p-value
Intercept	-0.543	0.037	$< 2e-16$
MeanSed	0.288	0.017	$< 2e-16$
MeanSed ²	-0.083	0.013	1.95e-10
EndProp	0.529	0.095	2.56e-08

Table 2.5.2: Fitted parameters for time-varying covariates in the GEE regression on average treatment

The fitted parameters are displayed in Table 2.5.2. In this table, the coefficient for MeanSed is presented after standardization. We refrain from interpreting these coefficient values here, as they specifically represent the performance of the current version

of SeqRTS. These values are not generalizable and fall outside the primary focus of our discussion. Our main intention is to demonstrate the directionality of the relationships: there is a positive correlation between average treatment and both MeanSed and EndProp, which aligns with our reasoning at the beginning of the section.

We now identify blocks with different performances and provide a recommendation for improving the average treatment to achieve 0.5 treatments per block for each participant. To identify blocks with different performances, we plot the relationship between average treatment and MeanSed. The plot is shown on the left in Figure 2.5.3. We controlled for EndProp by setting it at the 25th, 50th, and 75th percentiles (Q1, Q2, Q3) to make three fitted lines. It is clear that EndProp makes a difference, but is not too influential as the three fitted lines are close. We focus on the line setting EndProp to its 50th percentile.

A horizontal line representing the block average goal of 0.5 treatments is plotted. The points at which this line intersects with the fitted lines mark blocks where the algorithm’s performance deviates from its goal. A plausible categorization of the blocks in terms of different average treatments is as follows: $\text{MeanSed} < 10$, $10 \leq \text{MeanSed} < 20$, $20 \leq \text{MeanSed} < 30$, and $\text{MeanSed} \geq 30$. One strategy to improve the algorithm’s performance is to stratify blocks into these four categories. Before the next study batch, we could retrain the hyperparameters of the SeqRTS algorithm using the TUNE method described in [78] for these four types of blocks. During the study, before the start of each day, we can use prior K -day (e.g. prior 5-day) average of MeanSed as an input to determine the category of the block and then output the corresponding hyperparameters.

IDENTIFY TIME-VARYING COVARIATES TO IMPROVE UNIFORMITY

In this section, we seek ways to improve uniformity by first identifying time-varying covariates that are correlated with uniformity, and then provide a recommendation based on the regression result.

Note that MAD, the metric for uniformity, measures the differences among treatment

probabilities within a block. To select time-varying covariates that may correlate with MAD, we focus on the factors influencing difference between consecutive probabilities. Let t_k and t_{k+1} denote two consecutive risk times, where $t_{k+1} \geq t_k + 1$ ⁶. We consider the strong influencers on the ratio between consecutive probabilities, $\frac{p_{t_{k+1}}}{p_{t_k}}$, because when the ratio deviates from 1, variability in the whole set of probabilities in the block increases. From the SeqRTS formula (Equation 2.2), we observe that

$$p_{t_{k+1}} = \frac{N_{t_k} - p_{t_k}}{1 + \hat{g}_{t_{k+1}}} = \frac{N_{t_k} - N_{t_k}/(1 + \hat{g}_{t_k})}{1 + \hat{g}_{t_{k+1}}} = \frac{N_{t_k}}{(1 + \hat{g}_{t_k})} \frac{\hat{g}_{t_k}}{1 + \hat{g}_{t_{k+1}}} = p_{t_k} \frac{\hat{g}_{t_k}}{1 + \hat{g}_{t_{k+1}}}$$

where $N_t = \hat{N}_0 - \sum_{s=1}^{t-1} X_s \cdot p_s$ is remaining budget at time t (recall in Section 3 that a budget is an approximate average treatment allowed in the block). Moving p_{t_k} from right to the left, we have

$$\frac{p_{t_{k+1}}}{p_{t_k}} = \frac{\hat{g}_{t_k}}{1 + \hat{g}_{t_{k+1}}}. \quad (2.6)$$

Equation 2.6 indicates that the ratio between two consecutive treatment probabilities $\frac{p_{t_{k+1}}}{p_{t_k}}$ is characterized by the ratio between the estimated remaining sedentary times $\hat{g}_t$. Values of $\hat{g}_t$ shown in Figure 2.5.2 suggests two scenarios in which this ratio is large. First, when t_k and t_{k+1} are in the same hour, $\hat{g}_{t_k}$ is almost equal to $\hat{g}_{t_{k+1}}$. When they are equal, say $\hat{g}_{t_{k+1}} = \hat{g}_{t_k} = s$, the ratio in Equation 2.6 reduces to $\frac{s}{1+s}$. In this case, a smaller s leads to a larger ratio. Smaller s appears in the later hours of the block since $\hat{g}_t$ is smaller in the later hours of the block (as shown in Figure 2.5.2). Note that EndProp defined in Section 5.2.1 describes the proportion of risk times in the later hours. We therefore use EndProp as a regressor and hypothesize that it positively correlates with MAD. Second, when t_k and t_{k+1} are in different hours, Figure 2.5.2 shows that $\hat{g}_{t_k}$ differs largely from

⁶Here, $t_{k+1} \geq t_k + 1$ but not strictly equal to $t_k + 1$ to account for the fact that the next risk time after t_k might not be exactly the next decision time $t_k + 1$.

$\hat{g}_{t_{k+1}}$. As a result, both the ratio in Equation 2.6 and MAD should be high when risk times appear in different hours. To capture this phenomenon, we define HourVar (hour variation) as the number of distinct hours during which risk times occur in a block and use it as a regressor.

We now conduct a GEE regression analysis with EndProp and HourVar as two covariates and MAD as the outcome. An exploratory analysis indicates that EndProp, HourVar, HourVar², and EndProp $\times$ HourVar are correlated with the outcome. When fitting a GEE model with an AR-1 working correlation model, we found that HourVar² is not significant. We refine the GEE model to exclude HourVar². This refined model has Pearson residuals around zero, suggesting its plausibility. The coefficients for EndProp, HourVar, and their product are 2.20, 0.27, and -0.52, respectively, and all are significant.

With fitted regression model, we now identify blocks for which MAD is large. We plot the average MAD against EndProp, holding HourVar at its three quantiles. We use MAD = 0.01 as the benchmark to differentiate blocks with high and low MAD, as argued in Section 4. The plot is shown on the right of Figure 2.5.3. We observe the following from the plot: First, HourVar is not impactful since the three fitted lines look similar. Second, under the same HourVar, a large EndProp positively correlates with a large MAD.

These findings suggest that one way to improve uniformity is to gradually decrease $\hat{g}_t$ over decision times in a block. This approach aims to mitigate the impact of EndProp: as discussed above, the reason EndProp leads to high MAD is that $\hat{g}_t$ remains almost the same when two risk times are within the same hour. By decreasing $\hat{g}_t$, i.e., ensuring $\hat{g}_{t_{k+1}} < \hat{g}_{t_k}$ for two consecutive risk times t_k and t_{k+1} , instead of keeping them the same, the ratio $\frac{p_{t_k}}{p_{t_{k+1}}} = \frac{\hat{g}_{t_k}}{1+\hat{g}_{t_{k+1}}}$ will be closer to one, thereby reducing the MAD in the block. One simple strategy to achieve this is to estimate the proportion of sedentary times in a block (sedentary rate) and then let $\hat{g}_t$ be the product of the sedentary rate and the remaining decision times in the block. Other strategies might better improve uniformity, which we leave for future work.

2.6 CONCLUSIONS AND DISCUSSION

A new and challenging online decision problem has emerged in the evolving field of MRTs: determining a treatment probability that adheres to an average treatment constraint and a uniformity constraint when the times at which randomizations are allowed (risk times) are random. HeartSteps V2V3 MRT (HSV2V3) implemented the SeqRTS algorithm to meet these design constraints. Analyzing data from the HSV2V3, we found that SeqRTS performed promisingly in adhering to the core design constraints in the real world: the average treatment approximated the target of 1.5 per day, and the uniformity constraint was satisfied well on average across participants. We now discuss key considerations in evaluating, improving, and implementing the SeqRTS algorithm, in future studies.

2.6.1 HANDLING MISSING DATA IN THE EVALUATION

In the evaluation, data for the algorithm’s performance was missing at unavailable decision times and on unavailable days because the algorithm could not perform during those times and days. Below, we discuss how the missing data are handled in our analysis.

First, it is important to clarify both the assumptions and the target for the evaluation when excluding days on which a participant is unavailable. We have concluded that days we excluded were either days on which we are not interested in assessing the algorithm’s performance (days when a user is active or does not wear the Fitbit) or days on which we are interested (days when a user wears the Fitbit but there is no connection to the server), under which the data’s missingness is at random (MAR). As a result, the target for the evaluation is the algorithm’s performance on days when participants wear the watch and are not active all day, and the analysis using only the complete data does not pose a bias issue.

One potential concern of the result above is the influence of unobserved variables, such as mental tiredness (U), which may affect both a participant’s decision to wear the watch

and the accuracy of the algorithm’s predictors, like $\hat{g}_t$. As a result, the algorithm may seem to perform well by excluding days when participants do not wear the watch, potentially masking underlying challenges. We recognize that our analysis is based on a population representative of Kaiser Permanente patients newly diagnosed with stage 1 hypertension, and the algorithm’s performance is evaluated marginally across variables like U .

However, as patient populations evolve, the marginal effects may also change. Future studies should consider periodically updating the SeqRTS algorithm to ensure it remains effective in light of shifting behavior patterns.

Second, when handling unavailable decision times in a day, we assume that a participant does not respond to the anti-sedentary message, i.e., the anti-sedentary message does not affect a participant’s availability. This assumption is important for ensuring the correct interpretation of the evaluation results and the recommendations for the algorithm’s improvement, as otherwise, a participant’s responsiveness will entangle them. We illustrate this issue using the following example. Consider two users who differ in both sedentariness and responsiveness to the algorithm: User A is very sedentary and very responsive to messages. Being very sedentary means there is a high chance that one message is sent in a day, say 80%. Being very responsive means once they receive one message, they become active and not available for the rest of the day. This results in an average of 0.8 messages per day. User B is less sedentary but not responsive to the messages at all. Because User B is less sedentary, there is a lower chance of receiving one message in a day, say 50%. But because they do not respond to the message, there is also a chance, say 15%, that they receive two messages in a day. This results in an average of $1 \times 50\% + 2 \times 15\% = 0.8$ messages per day (assuming User B receives zero message on the remaining 35% of days). This example indicates that the algorithm’s performance on a given day could be impacted by both a user’s sedentariness and responsiveness. We assume users do not respond to the anti-sedentary message, making sedentariness the sole influencer of the algorithm’s performance. It will be interesting for future works to

explore methods to disentangle sedentariness and responsiveness.

2.6.2 THE METRIC FOR EVALUATING UNIFORMITY

Our key methodological contribution to assessing the algorithm’s performance is the proposal of mean absolute deviation (MAD) as a fairer metric for measuring uniformity than KL divergence, as proposed in [78]. This is because, in addition to measuring the algorithm’s performance on meeting the uniformity constraint, KL divergence also depends on its performance on the average treatment constraint. Poor performance on the average treatment constraint can result in a high KL value, even when the uniformity constraint is well met. In contrast, MAD measures SeqRTS’s performance on the uniformity constraint alone and is not influenced by the algorithm’s performance on the average treatment constraint.

2.6.3 WAYS TO IMPROVE THE ALGORITHM’S PERFORMANCE

With the ultimate goal of meeting two constraints for each participant, we have identified baseline and time-varying covariates that correlate with the algorithm’s performance and quantified their relationship to the algorithm’s performance. Based on the findings, we offer the following suggestions to improve the SeqRTS algorithm.

First, to improve uniformity, we should gradually decrease $\hat{g}_t$ over decision times in a block, instead of keeping $\hat{g}_t$ values nearly constant within the same decision hour, as done in HSV2V3. In Section 2.5.2, we showed that this “constant” pattern causes large variations in treatment probabilities for consecutive risk times within the same hour. One simple strategy to gradually decrease $\hat{g}_t$ is to estimate the proportion of sedentary times in a block (sedentary rate) and multiply it by the number of remaining decision times in the block.⁷ There could be other strategies that further improve uniformity. We leave

⁷This approach gradually decreases $\hat{g}_t$ because the sedentary rate is constant, but the remaining decision times in the block are deterministic and decrease with t .

this as a future work.

Second, to improve average treatment, we should use the degree of sedentariness to categorize participants and blocks and retrain the algorithm’s parameters for these categories. At the participant level, in Section 2.5.1, we found that participants who are highly sedentary during the baseline period are over-treated, those who are highly active during the baseline period are under-treated, and participants in the middle ground are well-treated. We suggest improving treatment averages by training three sets of hyperparameters for SeqRTS using data from these three types of participants. One way to train the hyperparameters is the TUNE method described in [78]. Once the hyperparameters are trained, one can use baseline variables in the next study to classify participants into the above three categories and assign them the corresponding hyperparameters to warm-start the study.

At the block level, in Section 2.5.2, we found that the prior 5-day average sedentariness (MeanSed) is positively correlated with the number of treatments. Specifically, blocks with $\text{MeanSed} < 10$ are under-treated. Blocks with $10 \leq \text{MeanSed} < 20$ are well-treated. Block with $20 \leq \text{MeanSed} < 30$ are over-treated, and blocks with $\text{MeanSed} \geq 30$ are too-overly-treated. As above, we could train the hyperparameters of the SeqRTS algorithm using TUNE for these four types of blocks. We recommend use MeanSed before the start of each block during the study as an input to determine the block’s category and then apply the appropriate hyperparameters. During the study, before the start of each block, the prior K -day (e.g., prior 5-day) average of MeanSed can be used as an input.

2.6.4 CONSIDERATIONS FOR REAL-WORLD IMPLEMENTATION AND DATA QUALITY IN SEQRTS

We conclude this case study with several considerations for the future implementation of the algorithm in the real world, aiming for easier and more accurate evaluation of the algorithm. First, code testing for the algorithm is crucial, as the cost of coding issues can

be substantial. During the evaluation phase, we encountered a coding issue that significantly impacted the evaluation process, as detailed in Appendix A.2. The issue resulted in the algorithm being unable to perform as intended (see Section 4.1 for a discussion). Although this issue affected a small proportion of the data, it had a considerable impact: including the affected data would improperly represent the algorithm’s performance on average treatment and uniformity. Consequently, we had to remove the affected data from the analysis, which was a loss.

Second, we need to enhance the quality of data collection during baseline period. In mobile health studies with MRTs, it is essential to consider post-study analysis when researchers collect data [129]. In Section 5.1, we highlighted that the significance of sedentariness and activity level variables collected from the baseline survey influence the algorithm’s performance, and we recommend researchers use them to improve the algorithm in future studies. However, the sedentariness data for 9 of the 82 participants was missing. Although we conducted sensitivity analyses using several imputation methods, it is always better when the data is not missing. Therefore, it is imperative to improve the quality of data collection concerning sedentariness and activity level at baseline in future studies.

Third, it is crucial to keep data used by the decision algorithm independent during a study. Specifically, in HSV2V3, we defined unavailability using four criteria but resorted to only one indicator when all availability criteria were met, instead of keeping four indicators, one for each availability criterion. The third unavailability criterion—unavailability within one hour after any type of notification is sent—created dependence for the sole availability indicator on the algorithm implemented in the study. As a result, future testing of different algorithms based on this study data will be unreliable. This issue is also pertinent when we find time-varying variables that correlate with the algorithm’s performance, as they all depend on the risk indicator, and hence the availability indicator. In the next iteration of SeqRTS, we suggest keeping four indicators, one for each availability

criterion. This suggestion should also apply to all stratified micro-randomized trials where risk times are of concern in decision making [35].

3

Chapter 2: Time-Varying Causal Survival Learning

3.1 INTRODUCTION

In healthcare, understanding the causal effect of medical interventions on patient survival is crucial. Heart transplantation is a compelling example, as demonstrated by the Stanford Heart Transplant study [30, 132], where patients with end-stage heart failure undergo surgery to replace their failing hearts with healthy donor hearts. While this procedure is likely to extend patients' lives on average, researchers are particularly interested

in how treatment effects vary with patient characteristics and surgical details. Time-to-event outcomes measure how long until a critical event occurs—in transplant cases, this is often the duration from surgery until death or organ rejection. Understanding these heterogeneous treatment effects—how the impact varies as a function of patient characteristics—is essential for improving patient selection criteria, optimizing intervention timing, and ultimately enhancing survival outcomes in transplant medicine [32, 64, 116]

However, measuring the causal effect of such medical interventions is not straightforward. After a patient is listed as a candidate for heart transplant, they must wait for an available donor heart before undergoing the procedure [10]. This means patients who receive heart transplants experience both control time (waiting period) and treatment time (post-transplant period). This differs from traditional causal inference where treatment and control groups are determined at the study’s outset and remain fixed throughout.

If there were a fixed time before which patients remained in the control group and after which they received treatment, then we could use that fixed time to separate treatment and control groups and conduct causal inference. However, this approach fails because treatment timing is random and highly variable. The waiting time for a heart transplant ranges from a few days to more than a year [40, 92]. The wide variation in treatment timing makes it impossible to establish any fixed time point that could meaningfully separate treatment and control groups [8].

Staggered adoption designs in econometrics [13] provide a way to estimate causal effects with random treatment times. These designs compare outcomes between treated and yet-to-be-treated units, using later-treated units as controls for earlier-treated ones. While these designs work well for repeatedly measured continuous outcomes, where a unit’s outcome is obtained at different times (e.g., to study a policy’s effect on a country’s GDP, you obtain the country’s GDP over many years, with some observations before the policy and some after), they cannot handle time-to-event outcomes that are observed only once—either at the event occurrence (such as mortality) or at study end.

To address this challenge, we extend staggered adoption designs to time-to-event outcomes by integrating survival analysis techniques [29, 43, 58, 68]. Specifically, we use hazard functions to model the instantaneous probability of an event. Hazard functions allow us to characterize each unit’s contribution to the outcome at any time, regardless of whether the unit is in its control or treatment period. We characterize a set of assumptions under which hazard-based models can be used to solve this causal inference problem.

In addition to handling random treatment timing, we address the complex, non-linear relationships between covariates and outcomes that often arise in real-world applications [50]. To handle these complexities, we employ double machine learning (DML) techniques [27, 45, 71, 90] in our estimation procedure. DML provides a powerful framework for improving estimation efficiency, allowing us to flexibly model non-linear relationships while maintaining robustness to potential model misspecifications.

The contributions of this paper are twofold. First, we bridge the gap between staggered adoption designs and survival analysis by identifying conditions under which time-varying treatment effects can be estimated in a survival framework. Specifically, we show how key assumptions from staggered adoption designs can be adapted to justify the use of existing survival analysis techniques for causal inference with time-varying treatments. Second, we propose an estimator that addresses the complexity of real-world data, enhancing performance through Double Machine Learning (DML) techniques to ensure unbiased and efficient estimation of treatment effects, thereby advancing survival analysis methods for handling time-varying treatments.

3.1.1 LITERATURE REVIEW

Our work builds on survival analysis, staggered adoption literature, and heterogeneous treatment effect literature. While survival analysis provides tools for modeling time-to-event data and recent causal inference literature offers insights into time-varying treat-

ments, the intersection—causal inference for survival outcomes with variable treatment timing—remains relatively unexplored. This gap needs attention because the problem is ubiquitous.

This methodological challenge extends beyond healthcare applications. In business settings, companies like streaming services and cloud storage providers need to understand how their interventions affect time-dependent outcomes. For example, when companies offer free membership trials, they must evaluate how free trial participation causally influences the time until conversion to paid memberships [9, 48]. As in transplant studies, both the timing of the intervention (free trial start) and the outcome (conversion) vary across units, making traditional causal inference methods insufficient for estimating these effects.

There have been several works combining causal inference with survival analysis. However, these approaches primarily focus on static treatments or simple time-varying confounders, rather than addressing the complexities of variable treatment timing. For example, [102] and [53] introduced key methodological frameworks for causal inference in survival settings. [75] developed doubly robust estimators for survival outcomes, while [130] proposed methods for handling time-dependent confounding. [119] considered time-varying treatments, but in their work, treatment timing is well-defined, such as scheduled half-year visits for people with HIV, and treatment can be reversed. In contrast, our work considers treatments with stochastic timing that cannot be reversed.

The challenges of treatments with random timing have been extensively studied in the econometrics literature, particularly in the context of difference-in-differences designs. Recent work by [13] and [47] has highlighted the importance of properly accounting for treatment timing in panel data settings. [111] provided crucial insights into the bias that can arise from ignoring treatment effect heterogeneity across adoption cohorts. [55] further developed these methods for handling staggered treatment adoption. However, despite its importance, this crucial structure of random treatment timing remains largely

unexplored in survival analysis settings, where our work pioneers new methodological approaches.

Our work employs Cox models for treatment modeling and focuses on learning the hazard ratio for survival outcomes, and this is different from other recent work on Cox models in causal inference. In related work, [107] developed a framework for testing treatment effects in time-varying settings using Cox models, but they concentrate on testing the null hypothesis of no treatment effect on continuous outcomes. In the biomedical domain, [131] considered time-varying treatments in electronic health records, but their setting differs from ours in that the treatment depends on time-varying covariates at each point, rather than following a staggered adoption pattern.

For heterogeneous treatment effect (HTE) estimation with time-to-event outcomes, researchers have developed various machine learning approaches, though none have explicitly addressed the staggered adoption design. [31] developed causal survival forests that adjust for right censored time-to-event outcome using doubly robust estimating equations, while [126] proposed censoring unbiased transformations with orthogonality properties that can be integrated with existing HTE learners in scenarios with or without competing risks. [127] provided a comprehensive review and tutorial on extending metalearners to handle right-censored time-to-event data. Taking a different approach, [45] applied the double machine learning framework to exponential families, with Cox models serving as a specific instance. More recently, [123] addressed estimation for left-truncated right-censored (LTRC) data, a significant advancement in the field. Despite these developments, the literature lacks methods specifically designed for causal inference in time-to-event settings with variable treatment timing as encountered in staggered adoption designs.

The organization of the paper is as follows: In Section 2, we formalize the notation, introduce assumptions for the causal framework, and present the statistical problem. In Section 3, we review existing methods for handling time-varying treatments in survival analysis. In Section 4, we introduce our double machine learning framework for robust

estimation of heterogeneous treatment effects. In Section 5, we present simulation results demonstrating the performance of our method. In Section 6, we analyze the Stanford Heart Transplant dataset to evaluate treatment effect heterogeneity. Section 7 concludes with a discussion of our findings and limitations.

3.2 PROBLEM SET UP

3.2.1 NOTATION AND DATA

Let capital letters denote random variables and lowercase letters denote their realizations. Consider $i = 1, \dots, n$ units. Let $T_i \in [0, \infty)$ denote the time until an event of interest occurs, such as time until mortality in transplant studies. Each unit i has a set of potential outcomes for T_i , denoted as $\{T_i(a) \in [0, \infty]\}$, where a represents the date (or time) when a binary treatment is first adopted by the unit. We refer to this as the adoption date, consistent with the terminology used in the staggered adoption literature (e.g., [13]). A unit can adopt the treatment at any of the time point $a \in [0, \infty)$, or not adopt the treatment at all during the time of observation, which we denote as $a = \infty$. We take a super-population perspective of $T_i(a)$, i.e., $T_i(a) \sim P$ are i.i.d. for some probability distribution P , the choice of which is discussed below. We observe for each unit in the population the adoption date $A_i \in [0, \infty]$. The observed event time of interest is denoted as T_i .

We also observe pre-treatment covariates $X_i \in \mathbb{R}^p$. We adapt the following standard causal assumptions from [106]:

Assumption 1 (Stable Unit Treatment Value Assumption, SUTVA): Each unit’s potential outcome is determined solely by its own treatment assignment, with no interference between units and uniform treatment versions. For each unit i ,

$$T_i = T_i(A_i),$$

where the observed outcome equals the potential outcome under the assigned treatment. This assumption tells how potential outcomes map to observed outcomes. This also assumes everyone receives the same version of treatment in healthcare settings, which might not hold in practice—for example, surgical procedures may vary by surgeon expertise or hospital resources—so caution is warranted when applying these methods in real-world clinical scenarios.

Assumption 2 (Unconfoundedness): The treatment assignment is unconfounded, conditional on covariates X_i . Formally,

$$A_i \perp\!\!\!\perp T_i(a) \mid X_i,$$

meaning that the treatment assignment A_i is independent of the potential outcome $T_i(a)$, given the covariates X_i .

In healthcare settings, this assumption may hold when treatment decisions are based solely on observed patient characteristics that are included in our covariate set. For heart transplantation, this would require that all factors influencing both transplant decisions and survival outcomes—such as disease severity, comorbidities, organ compatibility, and functional status—are measured and accounted for in our analysis. However, this assumption could be violated if unobserved factors like physician preferences, hospital protocols, or patient preferences that aren’t captured in medical records influence both transplantation decisions and survival outcomes.

Assumption 3 (Overlap): The probability of receiving treatment at time t , conditional on covariates, is strictly between 0 and 1 for all units. Specifically,

$$P(A_i \leq t \mid X_i) = a_t(X_i) \in [\epsilon, 1 - \epsilon] \quad \text{for some } \epsilon > 0.$$

This assumption ensures that each unit has a non-zero probability of receiving either

treatment or control. In healthcare systems, particularly for organ transplantation, this probability is often determined by allocation scores that rank patients based on observed covariates. For heart transplantation, factors such as blood type compatibility, tissue matching, disease severity, geographical proximity to donor, and waiting time all contribute to a patient’s position on the waiting list [25, 63, 95, 98]. This structured allocation system means that the overlap assumption requires careful validation, as certain combinations of covariates might effectively determine treatment timing with near certainty.

3.2.2 DISTRIBUTIONAL ASSUMPTION AND INTRODUCTION OF HAZARD

For the triplet of covariates, treatment, and outcome (X_i, A_i, T_i) , we impose the following general distributional assumptions:

$$\begin{aligned} X_i &\sim_{i.i.d.} f_X \\ A_i \mid X_i &\sim k(\cdot \mid X_i) \\ T_i(a) \mid X_i &\sim f(\cdot \mid a, X_i) \end{aligned}$$

Here, f_X represents the marginal density of the covariates on $\mathbb{R}^p$, without any additional parametric assumptions. The functions $k(\cdot \mid x)$ and $f(\cdot \mid a, X_i)$ denote conditional densities on $[0, \infty]$, corresponding to the treatment and outcome, respectively.

We also need to account for censoring. Censoring occurs when the event of interest—in this case, patient mortality—is not observed for all units within the study period. The censoring time, denoted as $C_i \in [0, \infty]$, represents the time at which the unit’s data becomes unavailable for observation.

Censoring can arise for several reasons: a patient may be lost to follow-up, the study period may end before mortality is observed, or administrative reasons may prevent fur-

ther observation. For censored units, the exact event time T_i is unknown; we only know that it exceeds the censoring time C_i . To handle this, we introduce a binary indicator variable Δ_i , where $\Delta_i = 0$ indicates censored data and $\Delta_i = 1$ indicates fully observed event times. Thus, the observed data consists of covariates X_i , treatment adoption time A_i , observed time $U_i = T_i \wedge C_i$ (the minimum of event time and censoring time), and censoring indicator Δ_i . We can represent each unit’s data as the tuple $(X_i, A_i, U_i, \Delta_i)$.

When censoring is present, we cannot directly estimate the distribution of event times. Instead of using the probability density function (pdf) $f(\cdot \mid a, X_i)$, we parameterize the distribution using the hazard function $h(\cdot \mid a, X_i)$ [29, 58]. The hazard function $h(t)$ at time t represents the instantaneous rate of event occurrence:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

The hazard function is particularly useful for censored data because it characterizes the risk of an event at time t , given survival up to that time. The relationship between the pdf and the hazard function is:

$$h(t) = \frac{f(t)}{1 - F(t)},$$

where $f(t)$ is the pdf and $F(t)$ is the cumulative distribution function (cdf). This relationship is one-to-one—specifying the hazard function uniquely determines the underlying distribution of the time-to-event variable [68].

Note that for each fixed x , the number of counterfactual hazard functions $h(\cdot \mid a, X_i)$ is infinite, as a is continuous on $[0, \infty]$. To address this complexity, we introduce two “exclusion” assumptions that simplify the model.

The first assumption states that the exact future transplant date doesn’t affect current outcomes [5, 6]:

Assumption 4: No Anticipation

For all i and for all adoption dates a such that $t < a$,

$$h(t \mid a, X_i) = h(t \mid \infty, X_i)$$

This reduces the infinite set of potential distributions for $t < a$ to a single one by assuming that before treatment adoption, the outcome event follows the baseline (or control) hazard f_0 . In practical terms, this means that future treatment adoption does not influence current outcomes.

The second assumption asserts that conditional on treatment adoption, the magnitude of the adoption time does not matter for potential outcomes, but only whether adoption has occurred by time t . This assumption is more restrictive but likely holds when a unit's characteristics X and event time T have a stable relationship that does not change with exposure duration. However, this assumption might not hold when X and T have a dynamic relationship—for example, when the effectiveness of the transplant depends on how long the patient has had it, or when patient characteristics change significantly over time post-transplant.

Assumption 5: Invariance to History

For all i and for all adoption dates a such that $t \geq a$,

$$h(t \mid a, X_i) = h(t \mid 0, X_i).$$

Together, assumption 4 and 5 enable us to simplify to only two hazard functions—one for the control, $h_0(t \mid X_i) = h(t \mid \infty, X_i)$, and one for the treated, $h_1(t \mid X_i) = h(t \mid 0, X_i)$. These assumptions have been widely adopted in the Difference-in-Differences (DID) literature. For a comprehensive list of related works, see Section 3.2 in [13].

We can now represent the hazard for the potential outcome, $h(t|a, X_i)$ by the following:

$$h(t|a, X_i) = h_0(t|X_i) \cdot \left(\frac{h_1(t|X_i)}{h_0(t|X_i)} \right)^{w_t}, \quad (3.1)$$

where $w_t := 1(t \geq a)$ is a binary function that indicates whether treatment has begun at time t .

We model the hazard using the proportional hazards model [29]:

$$\begin{aligned} h_0(t | x) &= \lambda(t) \exp(\eta_0(x)) \\ h_1(t | x) &= \lambda(t) \exp(\eta_1(x)), \end{aligned} \quad (3.2)$$

where $\lambda(t)$ represents the baseline hazard function, which captures the underlying risk of event at time t when all covariates are at their reference levels. Reference levels here denote the baseline values of covariates (zero for continuous variables or the specified baseline category for categorical variables) that define the baseline hazard to which all other covariate combinations are compared. The functions $\eta_0(x)$ and $\eta_1(x)$ represent the proportional hazards for the control and treatment groups. These functions can be either linear or non-linear functions of the covariates. A key advantage of this model is its ability to decouple the time component, $\lambda(t)$, from the covariate-dependent components, $\eta_0(x)$ and $\eta_1(x)$. This leads to a simplified treatment effect definition where the time component cancels out:

$$\begin{aligned} \tau(x) &= \log \left(\frac{h_1(t | x)}{h_0(t | x)} \right) \\ &= \log(h_1(t | x)) - \log(h_0(t | x)) \\ &= \eta_1(x) - \eta_0(x). \end{aligned} \quad (3.3)$$

We refer to this as the heterogeneous log hazard ratio (HLHR), which measures the

treatment effect on the hazard rate for a patient with covariate profile x .

In this paper, we assume that the treatment effect, $\tau(x)$, follows a linear parametric form. While more flexible specifications are possible, we focus on this linear specification for several reasons. First, in the context of heart transplant studies, key patient characteristics like age, medical history, and physiological measures often have approximately linear relationships with treatment outcomes [28]. Second, this specification mirrors the successful progression in the causal inference literature, where initial work on heterogeneous treatment effects for continuous outcomes began with linear models before expanding to more complex specifications [56, 62]. Specifically, for some $\beta \in \mathbb{R}^p$, we model $\tau(x)$ as:

$$\tau(x) = \beta^T x \tag{3.4}$$

As a result, the hazard function can be expressed as:

$$h(t \mid a, x) = \lambda(t) \exp(\eta_0(x) + w_t \cdot \tau(x)) \tag{3.5}$$

where $\lambda(t)$ is the baseline hazard and $w_t = 1(t \geq a)$ is an indicator function that denotes whether the treatment has been adopted by time t .

To estimate $\tau(x)$, we proceed in two steps. First, in Section 3.3, we review existing partial likelihood methods in survival analysis that perform well when $\eta_0(x)$ is correctly specified. These methods can achieve consistent estimation even when $\eta_0(x)$ belongs to a complex function space. Then, in Section 3.4, we introduce our main contribution: a double machine learning framework that provides robust estimation of heterogeneous treatment effects even when $\eta_0(x)$ is complex and potentially misspecified.

3.3 REVIEW OF HANDLING TIME-VARYING TREATMENT IN SURVIVAL MODELS

In this section, we review the incorporation of time-varying variables in survival models, as discussed in [42, 58].

3.3.1 REVIEW OF PARTIAL LIKELIHOOD AND ORDINARY COX REGRESSION

We begin by revisiting the ordinary Cox regression model to guide the reader through the derivation of maximum partial likelihood. In this subsection, we assume treatment is fixed from the start. Under this assumption, the hazard function takes the form:

$$h(t|w, x) = \lambda(t) \exp(\eta_0(x) + w \cdot \tau(x)) \quad (3.6)$$

where $\eta_0(x)$ represents the control group log hazard as a function of covariates x , and w is the treatment indicator.

The partial likelihood [29] is constructed by summing terms over the instances when an event (e.g., conversion) occurs, that is, when $\Delta_i = 1$ for a particular unit i . Let $\mathcal{R}_i = \{j : U_j \geq U_i\}$ denote the risk set for unit i , representing the set of individuals who have not yet experienced the event at time U_i . Furthermore, denote W_i as the treatment status for individual i . The log partial likelihood for the model is then expressed as follows:

$$\begin{aligned} \text{pl}_n(\tau, \eta_0) &:= \log \left(\prod_{\Delta_i=1} \frac{h(T_i|W_i, X_i)}{\sum_{j \in \mathcal{R}_i} h(T_i|W_j, X_j)} \right) \\ &= \sum_{\Delta_i=1} \left(\eta_0(X_i) + W_i \tau(X_i) - \log \left(\sum_{j \in \mathcal{R}_i} \exp(\eta_0(X_j) + W_j \tau(X_j)) \right) \right), \end{aligned} \quad (3.7)$$

To estimate the parameters η_0 and τ , we maximize this partial likelihood. This method

is widely used in practice due to its robustness, as it does not require explicit specification of the baseline hazard function $\lambda(t)$, while still maintaining desirable statistical properties [11].

3.3.2 REVIEW OF HANDLING TIME-VARYING TREATMENT

To accommodate time-varying treatment, we replace the fixed treatment indicator w with a time-varying indicator w_t :

$$h(t|a, x) = \lambda(t) \exp(\eta_0(x) + w_t \tau(x)) \quad (3.8)$$

where $w_t = 1(t \geq a)$ indicates whether the treatment has been initiated by time t .

To address the issue of time-varying covariates, we incorporate time variation into the partial likelihood framework [42, 58]. This extension allows for the correct modeling of covariates that change over time.

We retain the risk set $\mathcal{R}_i$ as defined previously. Let $W_i(t) = 1(A_i < t)$ represent the treatment status of unit i at time t . The partial likelihood for this time-varying model is then expressed as:

$$\begin{aligned} \text{pl}_n(\tau, \eta_0) := & \frac{1}{n} \sum_{\Delta_i=1} \left(\eta_0(X_i) + W_i(U_i) \tau(X_i) \right. \\ & \left. - \log \left(\sum_{j \in \mathcal{R}_i} \exp(\eta_0(X_j) + W_j(U_j) \tau(X_j)) \right) \right), \end{aligned} \quad (3.9)$$

This approach of incorporating time-varying treatments within the partial likelihood framework maintains desirable statistical properties and produces consistent estimators [58]. Recent work by [112] has extended this framework to allow lasso fitting when η_0 is non-linear.

While the above time-varying survival models provide a foundation for our work, they

face challenges in consistently estimating treatment effects when the baseline hazard is misspecified. To address this limitation, we now turn to a double machine learning framework that provides robustness to model misspecification.

3.4 DOUBLE MACHINE LEARNING ESTIMATOR

We now focus on efficient estimation of the treatment effect function $\tau(x)$ in the time-varying Cox proportional hazards model specified in Equation 3.6. As mentioned above, a straightforward approach would be jointly estimating $\tau(x)$ and the baseline hazard η_0 using the pseudo-likelihood in Equation 3.9.

However, this direct approach faces significant challenges because the baseline hazard η_0 acts as a nuisance function, and its estimation can interfere with the consistent estimation of the treatment effect. Traditional outcome-based methods that rely on correctly specifying the outcome model (in this case, the hazard function) are particularly vulnerable in this setting. This vulnerability arises because misspecification of the baseline hazard can directly bias the treatment effect estimates through the partial likelihood structure—errors in estimating η_0 propagate non-linearly through the risk set calculations, leading to biased estimates of $\tau(x)$.

To address these challenges, we adopt a double machine learning (DML) framework [27, 45, 71, 90]. The key insight of DML is to introduce propensity score estimation alongside the outcome model, providing double robustness and improved convergence rates.

Under the DML framework, the treatment effect estimator achieves favorable convergence rates through the product of nuisance parameter estimation errors. More precisely, let $e_0(x) = a_t(x)$ be our time-varying propensity score as defined in Equation 3.10, and let $\eta_0(x)$ be our baseline log hazard function as defined in Equation 3.6. If we denote the L2 convergence rates of their estimators as $\|\hat{e} - e_0\|_2 = O_p(r_n^e)$ and $\|\hat{\eta}_0 - \eta_0\|_2 = O_p(r_n^\eta)$

respectively, then through careful orthogonalization of the score function, the treatment effect estimator satisfies:

$$\|\hat{\tau} - \tau_0\|_2 = O_p(r_n^e \cdot r_n^\eta + \frac{1}{\sqrt{n}})$$

Here, r_n^e represents the rate at which our estimate of the propensity score converges to the true propensity score as sample size increases, while r_n^η represents the rate at which our estimate of the baseline log hazard converges to the true function. These rates capture the statistical efficiency of our nuisance parameter estimators, with smaller values indicating faster convergence.

This product structure is crucial: even if one nuisance component converges at a slower rate (e.g., $r_n^e = n^{-1/4}$), the treatment effect estimator can still achieve the optimal $\sqrt{n}$ -rate of convergence as long as the other component converges sufficiently fast (e.g., $r_n^\eta = n^{-1/4}$). This property, known as rate double robustness, makes the estimator robust to moderate misspecification of either nuisance component. For a comprehensive theoretical analysis of these convergence properties in the general framework of orthogonal statistical learning, we refer readers to [44].

3.4.1 TIME-VARYING CAUSAL SURVIVAL LEARNER (TV-CSL)

Building upon the work of [45], we propose TV-CSL (Time-Varying Causal Survival Learner) to handle time-varying treatments. The model is characterized by:

$$\begin{aligned} a_t(x) &= \frac{\int_0^t \mathbb{P}(\Delta = 1 \mid A = s, X) f(A = s \mid X) ds}{\int_0^\infty \mathbb{P}(\Delta = 1 \mid A = s, X) f(A = s \mid X) ds} \\ &= \mathbb{P}(A \leq t \mid \Delta = 1, X), \\ \nu_t(x) &= \tau(x) \cdot a_t(x) + \eta_0(x) \end{aligned} \tag{3.10}$$

Here, $a_t(x) = \mathbb{P}(A \leq t \mid \Delta = 1, X)$ represents the probability of adoption by time t for non-censored data. This is analogous to the “treatment probability” or the propensity

score at time t . When all data is not censored, $a_t(x) = \mathbb{P}(A \leq t \mid X)$. The full estimation procedure is presented in Algorithm 3.4.1.

Our work differs from [45], which developed a DML method for linear heterogeneous effects under a Cox model with treatment fixed at baseline, in two key aspects. First, the outcome models differ: their work uses the hazard form in Equation 3.6, while we use the time-varying form in Equation 3.8. Second, the propensity scores are distinct: their nuisance function maps covariates x to probabilities in $(0, 1)$, whereas our propensity score is a function of both x and t , representing the cumulative distribution of adoption time A conditional on X .

[t] [1] **Input:** Dataset $\{(X_i, T_i, \Delta_i, A_i)\}_{i=1}^n$, where X_i are covariates, T_i are survival times, Δ_i are event indicators, and A_i are treatment adoption dates

First Stage (Fold One): Estimate propensity score $a_t(x) = P(A_i \leq t \mid X_i = x, \Delta_i = 1)$

1) Estimate nuisance function ν_t by: 1. Maximizing the partial likelihood (Equation 3.9) to obtain $\hat{\eta}_0(x)$ and $\hat{\tau}(x)$ 2. Computing $\hat{\nu}_t(x) = \hat{\tau}(x) \cdot \hat{a}_t(x) + \hat{\eta}_0(x)$

Second Stage (Fold Two): Estimate treatment effect $\tau(x) = x^T \beta$ by solving:

$$\hat{\beta} = \min_{\beta'} \frac{1}{n} \sum_{\Delta_i=1} \left[\hat{\nu}_{\tau_i}(X_i) + (W_i(\tau_i) - \hat{a}_{\tau_i}(X_i)) X_i^\top \beta' \right. \\ \left. - \log \left(\sum_{l \in \mathcal{R}_i} \exp(\hat{\nu}_{\tau_i}(X_l) + (W_l(\tau_i) - \hat{a}_{\tau_i}(X_l)) X_l^\top \beta') \right) \right]$$

where: $W_i(t) = \mathbf{1}(A_i < t)$ Treatment status at time t $\mathcal{R}_i = \{j : U_j \geq U_i\}$ Risk set for subject i $\tau_i = U_i$ for i where $\Delta_i = 1$ Event times

Output: Estimated treatment effect function $\hat{\tau}(x) = x^T \hat{\beta}$

3.4.2 THEORETICAL JUSTIFICATION

Similar to existing causal inference literature [71, 90], we can derive theoretical results for the reduction of the learning rate.

Proposition 1 (Convergence Rate of Parameter Estimation) *Let the model for $a_t(x)$ be denoted as $\gamma(x)$. Under the following regularity conditions:*

1. *The covariates X are bounded, the true parameter β_0 lies in a bounded region $\mathcal{B}$, and the nuisance functions $\gamma_0(x)$, $\eta_0(x)$ along with their estimators $\gamma_n(x)$, $\eta_n(x)$ are uniformly bounded;*
2. *The minimal eigenvalues of the score derivative $\nabla_{\beta} s(\gamma(x), \eta(x), \beta)$ ¹ in $\mathcal{B}$ are lower bounded by some constant $C > 0$;*

If $\|\gamma_n(x) - \gamma_0(x)\|_2 = O(\alpha_n)$, $\|\eta_n(x) - \eta_0(x)\|_2 = O(\rho_n)$, and $\alpha_n \rightarrow 0$, $\rho_n \rightarrow 0$, then

$$\|\beta_n - \beta_0\|_2 = \tilde{O}(\alpha_n \rho_n + n^{-1/2}) \quad (3.11)$$

Proof: See Appendix B.3

Proposition 1 states that for $\hat{\tau}(x) = x^\top \hat{\beta}$ to reach a certain level of accuracy, the conditions on $\hat{a}_t(x)$ and $\hat{\eta}_0(x)$ are relatively loose. Specifically, if we want the estimate of β to achieve $n^{-1/2}$ convergence, we only need the product of the convergence rates of the outcome model $\eta_0(x)$ and treatment model $a_t(x)$ to be $n^{-1/2}$. For example, $\eta_0(x)$ could converge at rate $n^{-1/4}$ and $a_t(x)$ at rate $n^{-1/4}$. In comparison, outcome-based methods that do not use the treatment model $a_t(x)$ can only achieve $n^{-1/4}$ convergence rate.

This result build upon the framework established by [45]. It has profound implications for causal inference in survival settings with random treatment timing. The assumptions underlying this proposition require that our covariates and parameters remain within reasonable bounds, ensuring stability in our estimation procedure. The second assumption about minimal eigenvalues ensures that our estimation problem is well-conditioned and that small changes in the data do not lead to disproportionately large changes in our parameter estimates. Together, these assumptions create a framework where we can

¹See the Appendix B.3 for a definition of the score

achieve faster convergence by leveraging both the treatment and outcome models, even when each individual model might be estimated with moderate accuracy. This double robustness property is particularly valuable in healthcare applications like transplantation studies, where both the treatment assignment mechanism and the survival process are complex and difficult to model perfectly.

3.5 SIMULATION STUDY

In this section, we conduct comprehensive simulation studies to evaluate our method’s performance relative to existing approaches. We identify scenarios where our method demonstrates superior performance and conditions where its advantages are less pronounced, providing practical guidance for practitioners.

3.5.1 SIMULATION DESIGN

Let n denote the number of samples, and $X_i = (X_{i1}, X_{i2}, X_{i3})^\top \in \mathbb{R}^3$ denote the baseline covariates for unit i . $i = 1, \dots, n$, with sample sizes n varies among $\{200, 500, 1000, 2000\}$. We generate baseline covariates from a multivariate normal distribution: $X_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma = I_3$ is a 3×3 identity matrix. The treatment time A_i follows an exponential distribution with a rate parameter that depends linearly on covariates: $A_i | X_i \sim \text{Exp}(X_{i2} + X_{i3})$.

The survival time T_i is generated through a hazard function model $h(t | a, x) = \lambda(t) \cdot \exp(\eta_0(x) + 1(a \leq t)\tau(x))$ where $\lambda(t) = t$ is a linear baseline hazard function, $\eta_0(x)$ represents the baseline risk model, $\tau(x)$ captures the heterogeneous treatment effect, and $1(a \leq t)$ is the treatment indicator. Following [71], we specify the baseline risk model using a non-linear sigmoid-based function: $\eta_0(x) = -\frac{1}{2} \cdot \varsigma(X_1) \cdot \varsigma(X_{10})$ with scaled sigmoid $\varsigma(x) = \frac{2}{1 + e^{-12(x - \frac{1}{2})}}$. The treatment effect is specified as a linear combination of covariates: $\tau(x) = x_1 + x_2 + x_3$. We implement random censoring with a maximum follow-up time

of 20: $C_i = \min(20, \tilde{C}_i)$, where $\tilde{C}_i \sim \text{Exp}(0.1)$. These parameters ensure approximately 75% of observations are non-censored, with the censoring mechanism independent of both outcome and treatment. The simulation was repeated 100 times for each scenario, with sample sizes $n \in \{200, 500, 1000, 2000\}$ to evaluate the properties of the finite sample.

3.5.2 METHODS FOR COMPARISON

We evaluate two approaches for estimating heterogeneous treatment effects:

S-Lasso Method:² This method employs a single regression combining baseline risk η_0 and treatment effect τ through additive specification. It is "singly robust" as it does not incorporate the treatment model. For η_0 and τ specification, we consider two settings:

- Linear specification: $\eta_0(x) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ and $\tau(x) = \omega_0 W + \sum_{j=1}^3 \omega_j (W \cdot X_j)$
- Complex specification: Includes natural splines, squared terms, and all pairwise interactions. Specifically, $\eta_0(x)$ is linear in natural splines, squared terms, and all pairwise interactions of X , and $\tau(x)$ is linear in natural splines, squared terms, and all pairwise interactions of X multiplied by W . This results in a substantially larger feature space that captures non-linear relationships and interaction effects between predictors. While this more flexible specification can model complex relationships more accurately, it comes with increased risk of overfitting and reduced interpretability compared to the linear specification. The choice between these specifications represents a classic bias-variance tradeoff.

The combined model is fit using Lasso regularization with cross-validated penalty parameter.

TV-CSL Method: This is our proposed doubly robust estimation through cross-fitting in two stages:

²The name derives from "S-learner" as referenced in [71, 91], combined with our use of lasso regularization.

- First stage estimates the treatment model $a_t(x) = P(A \leq t \mid X = x)$ under both correct (all covariates) and misspecified (single covariate) settings
- Second stage estimates the baseline outcome model using Lasso with specifications matching S-Lasso

The Performance is evaluated using empirical Mean Squared Error (EMSE):

$$\text{EMSE} = \frac{1}{n} \sum_{i=1}^n (\hat{\tau}(X_i) - \tau(X_i))^2$$

where this calculation is performed on an independently generated testing set to evaluate out-of-sample performance. The MSE measures estimation quality by capturing both bias and variance components. All reported values are averages across 100 simulation replications.

3.5.3 RESULTS

Our simulation results demonstrate the relative performance of TV-CSL and S-Lasso under various specifications, focusing on estimation accuracy and robustness to model misspecification. We examine two key aspects: the impact of the treatment model specification and the performance under complex treatment effect specifications.

EFFECT OF TREATMENT MODEL SPECIFICATION

We first examine how the treatment model specification affects estimation quality. To isolate this effect, we maintain a correctly specified HTE model to ensure optimal conditions for both methods.

The results are shown in Figure 3.5.1. Both methods achieve lower MSE with the complex baseline outcome specification, which aligns with the true data-generating process where the baseline hazard (η_0) follows a non-linear pattern.

Comparing TV-CSL and S-Lasso, we observe that TV-CSL consistently outperforms S-Lasso regardless of treatment model specification, with the performance advantage being more pronounced when the treatment model is correctly specified. This aligns with our theoretical findings that the convergence rate depends on the product of errors in the nuisance estimators.

Notably, TV-CSL maintains its advantage over S-Lasso even when the treatment model is misspecified. This robustness can be attributed to the simplicity of our treatment model, where minor misspecifications have limited impact on the overall estimation error. The double machine learning framework effectively mitigates the impact of treatment model misspecification, allowing TV-CSL to maintain robust performance.

PERFORMANCE UNDER COMPLEX TREATMENT EFFECT SPECIFICATIONS

While our previous analysis focused on a linear (correctly specified) HTE model, we now evaluate the performance when using a complex model to estimate the HTE. Figure 3.5.2 presents these results.

For both methods, holding the outcome model fixed, the use of complex HTE specifications leads to higher MSE, though the magnitude of this increase varies between methods. This increased error can be attributed to the additional complexity in estimating the treatment effect model.

TV-CSL demonstrates superior performance relative to S-Lasso under two conditions: First, when the HTE model is correctly specified as linear, TV-CSL consistently outperforms S-Lasso across all sample sizes. This advantage stems from the double machine learning framework’s ability to reduce the impact of nuisance parameter estimation errors. Second, for complex HTE specifications, TV-CSL’s performance shows strong sample size dependency. While maintaining comparable performance at smaller sample sizes, TV-CSL outperforms S-Lasso at larger sample sizes.



Figure 3.5.1: Comparison of Mean Squared Error (MSE) between singly robust (S-Lasso) and doubly robust (TV-CSL) methods under varying conditions. Each panel shows a distinct combination of baseline hazard (η_0) and propensity score specifications, with rows representing η_0 complexity (linear/complex) and columns indicating propensity score specification (correctly specified/mis-specified). Results are shown across sample sizes (200, 500, 1000, 2000) on the x-axis, with MSE on the y-axis. Red and blue lines represent S-Lasso and TV-CSL performance, respectively. Error bars indicate ± 1.96 Monte Carlo standard errors.

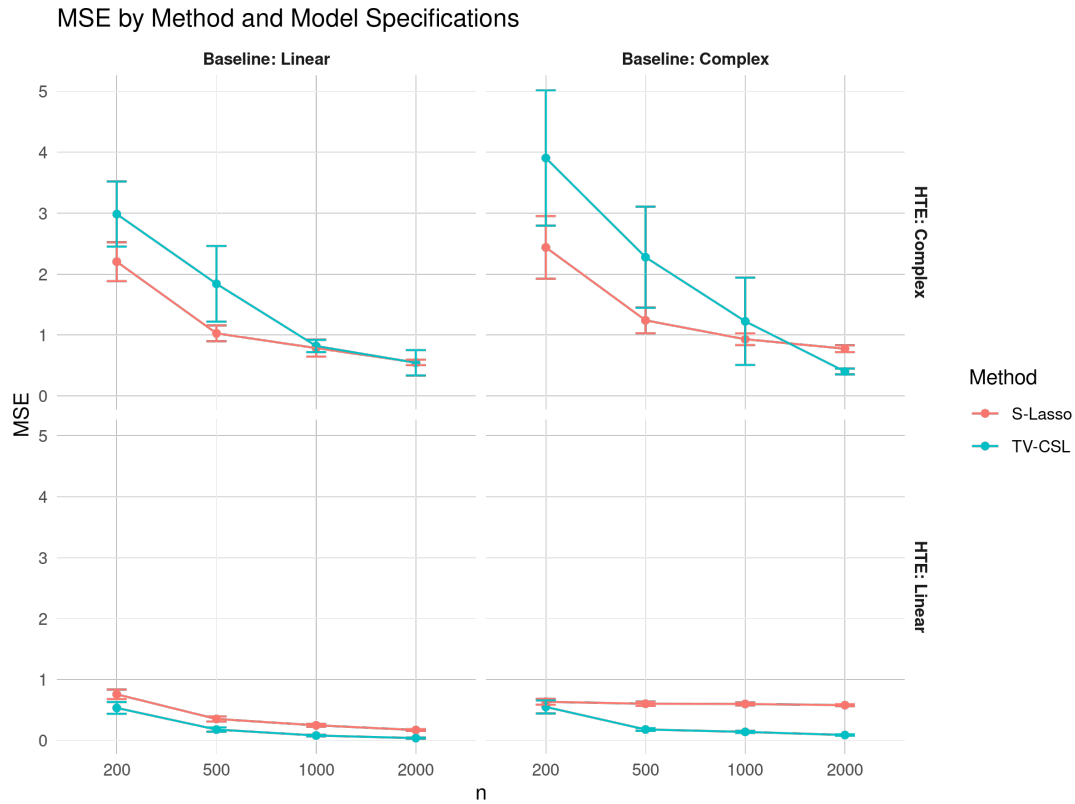


Figure 3.5.2: Comparison of Mean Squared Error (MSE) between S-Lasso and TV-CSL methods across different model specifications and sample sizes. Each panel represents a different combination of baseline hazard (η_0) and heterogeneous treatment effect (HTE) specifications, with rows indicating η_0 complexity (linear/complex) and columns indicating HTE complexity (linear/complex). The x-axis shows sample sizes (500, 1000, 2000), and the y-axis displays MSE. Red lines represent S-Lasso performance, while blue lines represent TV-CSL performance. Error bars indicate ± 1.96 Monte Carlo standard errors.

3.6 DATA ANALYSIS – STANFORD HEART TRANSPLANT DATASET

The Stanford Heart Transplant dataset originates from the pioneering efforts of the Stanford Heart Transplantation Program, which began in the 1960s. The program aimed to extend the lives of patients suffering from severe heart conditions by providing them with heart transplants. Patients were admitted based on strict medical criteria, and donor hearts were matched primarily by blood type. The dataset tracks patient survival times from program acceptance through three key phases: the initial enrollment period, the waiting period for a suitable donor heart, and the post-transplant period. The primary goal is to evaluate how heart transplants affect patient survival, while accounting for various patient characteristics such as age and previous surgical history [30].

In our analysis, we evaluate the performance of the TV-CSL method on the heart transplant dataset by examining: (a) the impact of incorporating time-varying information when evaluating transplant effects, and (b) the comparative effectiveness of TV-CSL with propensity score adjustments versus methods that do not use propensity scores.

3.6.1 SUMMARY STATISTICS

The study includes 103 participants. We have the following covariates: Age: Patient’s age at enrollment; Surgery: A binary indicator of whether the patient underwent surgery before or during the study; and Year: The time of enrollment, measured as years since the study’s initiation in 1967, capturing the evolution of medical practices and study conditions over time.

In our analysis, we set the start time (t_{start}) to 0 for all participants and include “Year” as a covariate representing time since study initiation. Following standard practice in survival analysis [68], we choose this approach rather than setting t_{start} to the value of “Year” and omitting it as a covariate. This ensures comparable risk assessment across patients because time zero represents a clinically meaningful baseline: the point at which

each patient was determined to be gravely ill and admitted to the study. Since all patients share this common clinical starting point, they have similar baseline hazards, aligning with the assumptions of the Cox proportional hazards model.

The summary statistics are presented in Table 3.6.1. In this study, the average age of participants is 45.17 years, with a standard deviation of 9.80 years. Approximately 16% of the participants had surgery before or during the study, as indicated by a mean value of 0.16 for the surgery variable. The year variable, representing time since the study began, has an average of 3.36 years with a standard deviation of 1.86 years, reflecting variability in enrollment timing among participants. Additionally, 67% of participants (69 individuals) received a heart transplant.

The distribution of heart transplants over time is depicted in Figure 3.6.1. This histogram illustrates the timing of transplants, highlighting any temporal trends or clustering of transplants throughout the study period. Understanding this distribution is essential for assessing time-related factors that may impact outcomes and for properly adjusting the model to account for these temporal effects.

	Variable	mean	sd
1	age	45.17	9.80
2	surgery	0.16	0.36
3	year	3.36	1.86
4	trt	0.67	0.47

Table 3.6.1: Summary Statistics of the Regressors (Mean and SD)

3.6.2 SHOWING EFFECT OF IGNORING IN TIME-VARIATION IN TREATMENT USING HTE

In this section, we assess the impact of incorporating time-varying information on the marginal effect of treatment. Specifically, we compare two Cox proportional hazards models: one that ignores the time-varying nature of the transplant variable and treats

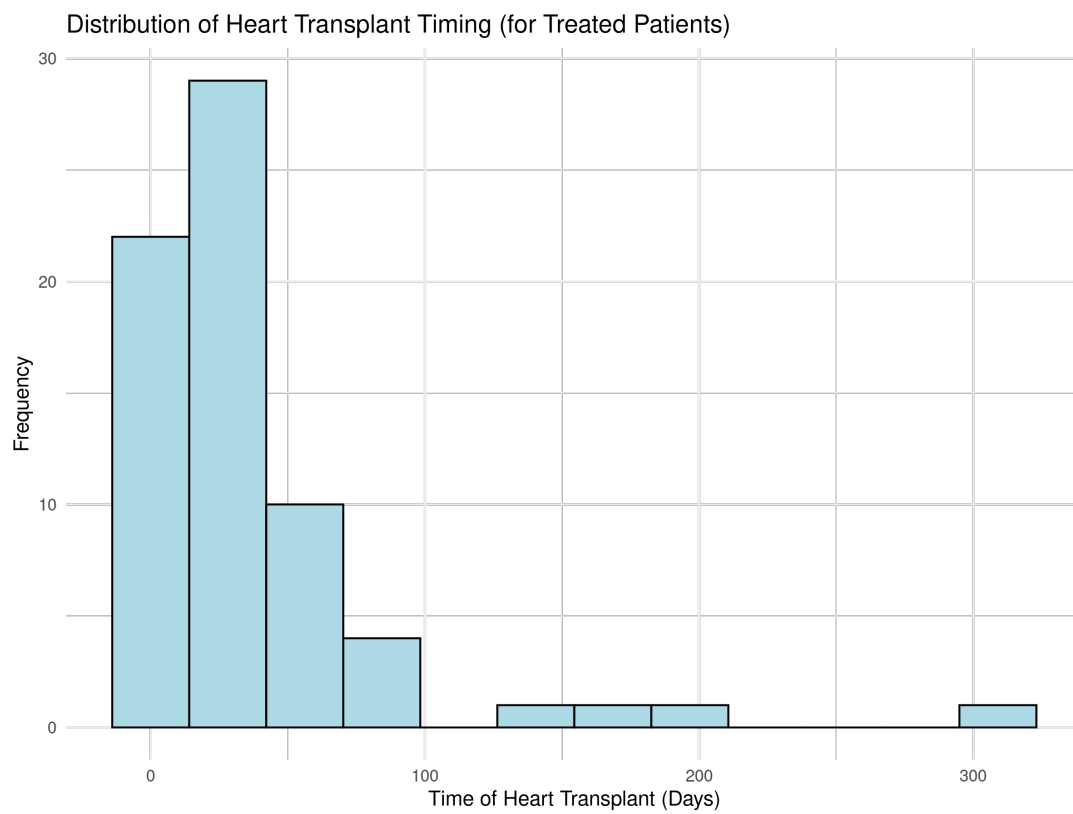


Figure 3.6.1: Distribution of Time of Heart Transplant

it as a fixed covariate, and another that includes the time-varying effect by dynamically updating transplant status throughout the study.

Here we consider the same baseline regressors (age, surgery, and year), but now use them to estimate the heterogeneous treatment effects (HTE). We focus on the coefficients related to the interaction between the treatment (transplant status) and the baseline covariates to understand how the treatment effect varies across different subgroups of the population.

Table ?? shows differences between the two models. In the model ignoring treatment time, the surgery-treatment interaction shows a significant effect (coef = -2.191 , p-value < 0.01), suggesting a transplant benefit for patients who have had previous surgery. However, this effect disappears in the model that includes treatment time (coef = -0.557 , p-value = 0.47). The treatment variable itself also shows this difference. Other interactions remain non-significant in both models.

To summarize, ignoring time variation can lead to inflated treatment effect estimates. This occurs because ignoring time variation mistakenly treats control time as treatment time, whereas accounting for time variation correctly captures the temporal nature of treatment administration.

2*Variable	Ignore Treatment Time		Include Treatment Time	
	Coef (SE)	P-value	Coef (SE)	P-value
Trt	-1.504 (0.292)	0.00	0.117 (0.340)	0.73
Age $\times$ Trt	-0.259 (0.285)	0.36	0.286 (0.254)	0.26
Surgery $\times$ Trt	-2.191 (0.778)	0.00	-0.557 (0.777)	0.47
Year $\times$ Trt	0.206 (0.261)	0.43	0.421 (0.260)	0.11

Table 3.6.2: Comparison of Cox Models: Fixed vs. Time-varying

3.6.3 EFFECTS OF MACHINE LEARNING AND TIME-VARYING CAUSAL SURVIVAL LEARNING

After demonstrating the effect of ignoring time-variation in treatment, we now focus on including time-variation and comparing an outcome-based method (S-lasso) to a cross-fit, doubly robust method (TV-CSL).

To do this, we first estimate η_0 and $\tau(x)$ from the original dataset using TV-CSL and use these estimates as ground truth. To achieve reliable estimates, we exclude the binary variable surgery because it has a true-to-false ratio of 76% (82 of 108). During cross-fitting, this variable could result in even more extreme ratios, hence leading to unreliable estimates. We also use a linear model for η_0 because, as observed in simulations, simpler models are preferable with small sample sizes (the sample size here is 108) to avoid issues with cross-fitting.

We then design a linear treatment model $a(x)$ and sample the treatment times. Note that these treatment times represent the waiting period after a patient enters the registry and is primarily matched by blood type (though blood type data is unavailable). Age should not be a dependent factor, but years of admission could impact treatment time to model donor availability. Therefore, we use a univariate linear model: $A|X \sim \text{Exp}(\alpha_0 + \alpha_1 \text{Years})$.

We repeat the sampling-estimation procedure 100 times. The comparison is based on the average mean squared difference between the estimated effect and the "true" effect.

Table 3.6.3: MSE by Method and Baseline Outcome Model

Method	Baseline Outcome Model	
	complex	linear
S-lasso	0.386	0.492
TV-CSL	1.220	1.150

Table 3.6.3 presents the mean squared error (MSE) comparison between different methodological approaches. We have the following findings: First, TV-CSL performs worse than the single-fit method across both baseline models. This aligns with our simulation findings regarding small sample performance. Second, when using single S-lasso, the complex baseline outcome model shows modest improvement over its linear counterpart, with MSEs of 0.386 and 0.492 respectively.

3.7 CONCLUSION

In this paper, we propose a novel framework for estimating causal effects of time-varying treatments on time-to-event outcomes by extending the staggered adoption framework from econometrics to a survival analysis setting. Our approach leverages the Cox proportional hazards model and incorporates double machine learning (DML) to address complexities in real-world data, such as nonlinear covariate relationships and high-dimensional settings. Through simulations, we demonstrate that our estimator effectively reduces bias and improves efficiency compared to traditional outcome-based methods.

The implications of this work extend beyond healthcare to domains with time-dependent outcomes and random intervention timing. In healthcare, our method enables better understanding of how interventions like organ transplants affect patient survival based on individual characteristics, helping improve patient selection and timing of procedures. In business settings, particularly subscription-based services, our framework enables more accurate estimation of how interventions like free trials causally affect time-to-conversion outcomes. By accounting for both intervention timing and unit characteristics while maintaining methodological rigor, our approach provides reliable insights for optimizing customer acquisition and retention strategies across industries.

Our proposed estimator advances the capabilities of causal inference in survival analysis, providing a robust approach for analyzing staggered treatment adoption with time-varying

interventions in diverse applied contexts.

3.7.1 LIMITATIONS AND FUTURE WORK

Our study has several limitations that warrant discussion. A key limitation is our assumption of linear treatment effect heterogeneity. Although this specification allows substantial methodological progress and provides interpretable results for medical decision-making, treatment effects in complex medical interventions may exhibit non-linear patterns across patient characteristics. Future work could extend our framework to accommodate more flexible specifications of $\tau(x)$ using non-parametric or semi-parametric approaches.

Another limitation is that in reality, there is an instantaneous increase in risk immediately after heart transplant. This occurs because some patients may experience severe rejection reactions when receiving a new heart [81]. After surviving this critical period, the patient’s risk typically decreases. In future work, rather than modeling a single hazard function post-transplant as we did, we should consider two distinct hazard functions: one capturing the elevated risk immediately after transplant, and another reflecting the lower risk level that follows successful adaptation.

Another promising direction is to develop testing procedures for identifying which patients benefit most from transplantation, building upon our effect estimation framework. While [38] established testing procedures for continuous outcomes, similar methodologies could be developed for time-to-event outcomes.

3.8 INTRODUCTION

Matching and weighting estimators are fundamental tools in causal inference for estimating treatment effects. Matching methods pair treated units with similar control units based on observed covariates [104, 105] while weighting approaches adjust for confounding by reweighting observations to achieve covariate balance [54, 57]. These methods have been widely adopted across diverse fields, including economics [33, 51], epidemiology [110], and policy evaluation [108].

[1] established the foundational asymptotic theory for matching estimators, revealing their nonstandard behavior due to the fixed number of matches. Unlike other nonparametric treatment effect estimators [52, 54], matching estimators are generally not $\sqrt{N}$ -consistent. This discovery led to important developments in bias correction [3] and inference methods. Notably, [2] demonstrated that the standard bootstrap fails to provide valid inference for matching estimators, recommending instead the use of analytical standard errors or subsampling methods [96].

[93] proposed a wild bootstrap procedure that is theoretically valid under certain asymptotic conditions. However, our empirical investigations reveal that this procedure can produce unreliable inference when there is substantial overlap in matched samples—a common scenario when the number of treated units is small but the number of control units is large. Here, by overlap, we mean the phenomenon where matched controls for different treated units contain many common control units; that is, the same control units appear repeatedly in the matched sets of multiple treated units. This overlap creates complex dependencies that are not properly accounted for in existing inference procedures [4]. This dependency structure becomes particularly problematic when constructing confidence intervals using bootstrap methods, as the resampling scheme fails to capture the true variance of the estimator.

We address this challenge through two main contributions. First, we analyze an alter-

native inference procedure that demonstrates robust performance in simulations when the wild bootstrap fails to produce valid confidence intervals under substantial overlap conditions. This procedure is a Wald-type confidence interval based on a variance estimator that has been previously used in empirical work (e.g., [26, 61]), but whose comparative advantage over state-of-the-art methods like [93] has not been systematically established. Our comprehensive comparison demonstrates that this estimator maintains validity even under substantial overlap in matched samples, outperforming existing approaches. Second, we develop a theoretical framework that rigorously justifies the estimator’s validity by establishing its consistency and asymptotic normality, providing practitioners with a sound statistical foundation for its application.

Our contributions are as follows:

- **Theoretical Analysis of a Practical Variance Estimator:** We provide a rigorous theoretical analysis of a variance estimator previously proposed in [26]. Our analysis demonstrates that this estimator is computationally efficient, theoretically justified, and more practical than alternatives. Unlike the estimator in [1] which requires matching within both treatment groups, this approach only requires matching treated units to controls, making it particularly valuable for effect estimation with small treated samples. Furthermore, we find that this estimator can outperform current state-of-the-art alternatives in settings that mimic real-world matching contexts. In particular, our estimator remains robust when control units are reused as matches for multiple treated units. Through carefully designed simulation studies, we show that it maintains proper coverage while the wild bootstrap method proposed by [93], which is the current state-of-the-art method for inference in matching, fails in these common scenarios. We also establish that the estimator possesses heteroskedasticity-consistent properties, drawing important parallels to the well-known Huber-White [125] robust standard errors in regression analysis. This con-

nection creates a theoretical bridge between matching-based and regression-based inference methods, unifying seemingly disparate approaches to variance estimation in causal inference.

- **Generalized Theoretical Framework:** While [26] use this approach in the context of radial matching, we show that our framework applies to a variety of matching contexts including nearest neighbor matching, propensity score matching, and synthetic control weighting. We introduce two novel conditions to the matching literature that together create a more practical framework for valid inference in matching estimators. First, our derivative control condition relaxes traditional requirements about how outcome functions can change. While previous work by [1] required the outcome function to change at a constant rate everywhere (Lipschitz continuity), our approach allows the rate of change to vary across the covariate space as long as it is appropriately balanced by the size of matched clusters. This innovation permits valid inference even with functions that have steep gradients in some regions. Second, our shrinking clusters assumption does not specify how quickly matches must improve with sample size, only that they eventually become arbitrarily close. This flexibility, new to the matching literature, accommodates many matching methods beyond the specific M -NN approach in earlier work, including radius matching and propensity score techniques. Together, these conditions significantly expand the applicability of matching methods while ensuring theoretical guarantees, allowing researchers to apply these techniques across a wider range of real-world research scenarios with greater confidence in their statistical properties.

- **Extended Applications to Other Causal Inference Estimators:** We demonstrate that the variance estimation framework extends beyond matching to other causal inference methods, particularly weighting estimators [134?]. Through additional simulations with challenging datasets, we show that our variance estimation

approach maintains proper coverage when applied to stable balancing weights, suggesting potential for creating a more coherent inference framework for both matching and weighting approaches in causal inference.

3.9 PROBLEM SETUP

3.9.1 MODEL

We consider a setting with n observations, each representing a unit in our study population. The sample consists of n_T treated units and n_C control units, with $n = n_T + n_C$.

For each unit i , we observe a tuple $\{Z_i, Y_i, \mathbf{X}_i\}$ where:

- $Z_i \in \{0, 1\}$ denotes its binary treatment status.
- $Y_i \in \mathbb{R}$ denotes its observed real-valued outcome.
- $\mathbf{X}_i \equiv \{X_{1i}, \dots, X_{ki}\}^T \in \mathbb{R}^k$ denotes its k -dimensional real-valued covariate vector.

We adopt the potential outcomes framework where each unit has two potential outcomes: $Y_i(1)$ and $Y_i(0)$. Here, $Y_i(1)$ represents the outcome if unit i receives treatment, and $Y_i(0)$ represents the outcome if unit i does not receive treatment. The fundamental problem of causal inference is that we only observe one of these potential outcomes for each unit. Specifically, the observed outcome for unit i is $Y_i \equiv (1 - Z_i)Y_i(0) + Z_iY_i(1)$ under the stable unit treatment value assumption (SUTVA).

We assume the data generating process follows independent and identically distributed (i.i.d.) sampling of the potential outcome tuples $\{Y_i(0), Y_i(1), Z_i, \mathbf{X}_i\}_{i=1}^n$. For each unit i , the generic random variables $(Y(0), Y(1), Z, \mathbf{X})$ represent the population distribution from which the observed data are drawn. Throughout the paper, indexed variables (e.g., $\mathbf{X}_i$) refer to specific observations, while non-indexed variables (e.g., $\mathbf{X}$) refer to the generic random variables representing the population distribution.

To proceed with estimation, we make the following assumptions:

Assumption 1 (Compact support) *The covariate vector $\mathbf{X}$ is a k -dimensional random vector with components that are continuous random variables, distributed on $\mathbb{R}^k$ with compact support $\mathbb{X}$. The density of X is bounded and bounded away from zero on its support.*

The compact support assumption helps ensure that the covariate space is well-behaved, which facilitates consistent estimation and rules out pathological cases where the distribution of covariates becomes too sparse or unbounded.

Assumption 2 (Unconfoundedness and overlap [106]) *For almost every $x \in \mathbb{X}$:*

1. $(Y(1), Y(0)) \perp\!\!\!\perp Z \mid \mathbf{X}$
2. $\Pr(Z = 1 \mid X = x) < 1 - \eta$ for some $\eta > 0$

This assumption states that, conditional on the observed covariates, treatment assignment is independent of the potential outcomes, and that both treated and control units are sufficiently represented across the covariate space.

Importantly, the treatment indicators Z_i are randomly drawn according to the treatment assignment mechanism, which implies that the number of treated units n_T is a random quantity even when the total sample size n is fixed. This leads to our next assumption:

Assumption 3 (Sampling) *Conditional on $Z_i = z$, the sample consists of independent draws from the distribution of $(Y, X \mid Z = z)$ for $z \in \{0, 1\}$. As the sample size $n \rightarrow \infty$, we have $n_T^r/n_C \rightarrow \theta$ for some $r \geq 1$ and $0 < \theta < \infty$.*

We further assume a model where potential outcomes are generated as:

$$Y_i(0) = f_0(\mathbf{X}_i) + \epsilon_{0,i}$$

$$Y_i(1) = f_1(\mathbf{X}_i) + \epsilon_{1,i}$$

Here, $f_0(\mathbf{X}) \equiv E[Y(0)|\mathbf{X}]$ and $f_1(\mathbf{X}) \equiv E[Y(1)|\mathbf{X}]$ are the true conditional expectation functions of the potential outcomes under control and treatment, respectively (often referred to as “response surfaces” in the causal inference literature [? ?]). The error terms $\epsilon_{0,i}$ and $\epsilon_{1,i}$ represent the deviations of the individual potential outcomes from their respective conditional expectations, with conditional variances $\sigma_{0,i}^2$ and $\sigma_{1,i}^2$ respectively. Further distributional assumptions about these error terms are detailed in Section 3.10.2.

3.9.2 ESTIMAND AND THE ESTIMATOR

We define $\tau(\mathbf{X}_i) = f_1(\mathbf{X}_i) - f_0(\mathbf{X}_i)$ as the systematic treatment effect, which represents the systematic component of the treatment effect for units with covariates $\mathbf{X}_i$. Note that the individual treatment effect $Y_i(1) - Y_i(0) = \tau(\mathbf{X}_i) + (\epsilon_{1,i} - \epsilon_{0,i})$ includes both this systematic component and an idiosyncratic component. We consider the estimand to be the population average treatment effect on the treated (ATT)

$$\tau = E[f_1(X_i) - f_0(X_i) \mid Z_i = 1],$$

We write the set of all treated units’ indices as $\mathcal{T} = \{i : Z_i = 1\}$, the set of all control units’ indices as $\mathcal{C} = \{i : Z_i = 0\}$, and $t \in \mathcal{T}$, $j \in \mathcal{C}$ as individual treated and control units respectively. We denote the set of indices of control units matched to a treated unit $t \in \mathcal{T}$ as $\mathcal{C}_t = \{j \in \mathcal{C} : \text{unit } j \text{ is matched to unit } t\}$, which is determined by a matching procedure that maps the observed data to these match sets. Finally, we denote the size of a set $\mathcal{S}$ as $|\mathcal{S}|$.

A matching procedure pairs each treated unit with one or more control units that have similar covariate values, thus approximating the counterfactual outcome for the treated unit. Given such a matching procedure, we define the matching estimators for the ATT:

$$\hat{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \sum_{j \in \mathcal{C}_t} w_{jt} Y_j \right). \quad (3.12)$$

where $w_{jt} \in [0, 1]$ is the weight assigned to the matched control unit j for treated unit t , with $\sum_{j \in \mathcal{C}_t} w_{jt} = 1$ for each $t \in \mathcal{T}$. For example, in M -nearest neighbor (M -NN) matching [1, 105, 110], $\mathcal{C}_t$ consists of the closest M neighbors to unit t based on covariate distance, and each neighbor receives equal weight $w_{jt} = 1/M$. Another example is the synthetic control approach in [26], which first obtains $\mathcal{C}_t$ local radius matching, and then determines the weights w_{jt} by solving a convex optimization problem that minimizes the distance between the treated unit's covariates $\mathbf{X}_t$ and the weighted average of control units' covariates $\sum_{j \in \mathcal{C}_t} w_{jt} \mathbf{X}_j$.

Define the matching radius for a treated unit t with covariate value $\mathbf{X}_t$ as:

$$r(\mathcal{C}_t) = \sup_{j \in \mathcal{C}_t} \|\mathbf{X}_t - \mathbf{X}_j\|.$$

This radius represents the maximum distance between a treated unit and any of its matched controls. The probabilistic properties of this radius will be crucial for establishing our theoretical results.

Assumption 4 (Exponential Tail Condition) *The matching radius satisfies:*

$$P \left(n_C^{1/k} r(\mathcal{C}_t) > u \right) \leq C_1 \exp(-C_2 u^k),$$

where C_1, C_2 are positive constants, k is the dimension of the covariate space, and M is the number of matches.

This assumption requires that the probability of having a large (scaled) matching radius decays exponentially, which ensures that the matches become increasingly accurate as the sample size grows. This is a more precise characterization than simply requiring that clusters shrink asymptotically, as it specifies the rate at which the tail probability diminishes.

Several common matching methods satisfy the exponential tail condition under appropriate implementation, including M -NN matching and radius matching. When using fixed M -NN matching, the exponential tail condition is satisfied provided that the density of covariates is bounded and has overlapped support, as established by [1]. In this approach, each treated unit is matched to its M closest control units based on covariate distance, with each neighbor receiving equal weight $w_{jt} = 1/M$.

Another approach is radius matching, where we set the matching radius as $D(n_C) = cn_C^{-1/k}$. To provide an intuitive understanding of this choice, we can use the following heuristic argument. This choice ensures two important properties. First, the probability of obtaining at least one match for a treated unit approaches 1 as $n_C \rightarrow \infty$. Heuristically, the expected number of controls in a ball of radius $D(n_C)$ around a treated unit with covariates $\mathbf{X}_t$ is approximately

$$n_C \cdot f(\mathbf{X}_t) v_k [cn_C^{-1/k}]^k \approx f(\mathbf{X}_t) v_k c^k,$$

where v_k is the volume of the unit ball in $\mathbb{R}^k$ and $f(\mathbf{X}_t)$ is the density at $\mathbf{X}_t$.

Second, this matching scheme has an exponential tail for the scaled discrepancy $n_C^{1/k} \|\mathbf{X}_j - \mathbf{X}_t\|$. By analyzing the order statistics of nearest neighbor distances and applying large-deviation bounds, we can show that

$$\Pr(n_C^{1/k} \|\mathbf{X}_j - \mathbf{X}_t\| > u) \leq C_1 \exp(-C_2 u^k)$$

for some constants $C_1, C_2 > 0$, thus satisfying the exponential tail condition.

In the next section, we address the inference problem, focusing on the asymptotic normality of the matching estimator and the decomposition of its variance components. This provides the foundation for constructing valid confidence intervals. Following this, we turn our attention to the crucial challenge of variance estimation. We introduce a consistent estimator that accounts for both homoskedastic and heteroskedastic error structures, refining previous approaches to improve efficiency and robustness.

3.10 THE INFERENCE PROBLEM

To construct valid confidence intervals for our matching estimator $\hat{\tau}$, we require asymptotic normality of the form:

$$\frac{\sqrt{n_T}(\hat{\tau} - \tau)}{V^{-1/2}} \xrightarrow{d} N(0, 1).$$

The difference between the matching estimator $\hat{\tau}$ and the estimand τ can be decomposed into three components:

$$\hat{\tau} - \tau = \hat{\tau} - \tau_{\text{SATT}} + \tau_{\text{SATT}} - \tau = B_n + E_n + P_n \tag{3.13}$$

where

$$B_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{C}_t} w_{jt} (f_0(X_t) - f_0(X_j))$$

represents bias from imperfect covariate matching.

$$\begin{aligned} E_n &= \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(\epsilon_t - \sum_{j \in \mathcal{C}_t} w_{jt} \epsilon_j \right) \\ &= \frac{1}{n_T} \sum_{t \in \mathcal{T}} \epsilon_t - \frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j \epsilon_j \end{aligned}$$

captures measurement error from random variation in unobserved factors.

$$P_n = \tau_{\text{SATT}} - \tau$$

measures population error between sample and population treatment effects.

where τ_{SATT} is the sample average treatment effect on the treated (SATT):

$$\tau_{\text{SATT}} = \frac{1}{n_T} \sum_{t \in \mathcal{T}} (f_1(X_t) - f_0(X_t)).$$

We now analyze the key components of our inference framework in detail. Section 3.1 examines the bias term and its asymptotic behavior under different matching schemes. Section 3.2 introduces critical assumptions about error variance that underpin our theoretical results. Section 3.3 develops the variance decomposition, separating contributions from measurement error and population heterogeneity. Together, these elements establish the foundation for our central limit theorem.

3.10.1 THE BIAS TERM AND ITS CONVERGENCE RATE

A crucial challenge in establishing the asymptotic normality of matching estimators is the slow convergence rate of the bias term. Following [1], under regularity conditions on data distribution introduced at Section 3.9.1, this bias term converges at a rate of $O_p(n_T^{-1/k})$,

where k is the dimension of the covariate space. This rate is typically slower than the $n_T^{-1/2}$ rate required for standard asymptotic normality results.

Proposition 2 (Bias convergence rate) *Under Assumptions 1, 2, and 3, if $f_0(x)$ is Lipschitz continuous on $\mathbb{X}$, then*

$$Bias = O_p(n_T^{-1/k})$$

This slow convergence rate of the bias term necessitates explicit bias correction for valid inference. While various approaches to bias correction exist, including the method proposed by [3], our focus in this paper will be on variance estimation conditional on a bias correction procedure.

3.10.2 ERROR VARIANCE ASSUMPTIONS

To analyze the large-sample behavior of our variance estimator, we place structure on the conditional variance of the potential outcomes. This section introduces the regularity conditions we require on the variance functions $\sigma_0^2(x)$ and $\sigma_1^2(x)$.

Let us denote the conditional variances of the potential outcomes as:

$$\begin{aligned}\sigma_{0,i}^2 &= E[(Y_i(0) - f_0(\mathbf{X}_i))^2 | \mathbf{X}_i] = E[\epsilon_{0,i}^2 | \mathbf{X}_i], \\ \sigma_{1,i}^2 &= E[(Y_i(1) - f_1(\mathbf{X}_i))^2 | \mathbf{X}_i] = E[\epsilon_{1,i}^2 | \mathbf{X}_i].\end{aligned}\tag{3.14}$$

We now define a class of variance functions with properties that enable consistent estimation in the matched setting.

[Regular variance function] A function $\sigma^2 : \mathcal{X} \rightarrow \mathbb{R}_+$ is said to be a *regular variance function* if it satisfies the following:

- **Uniform continuity.** $\sigma^2(\cdot)$ is uniformly continuous (or Lipschitz) on the support $\mathcal{X} \subset \mathbb{R}^d$ of X .

- **Boundedness.** There exist constants $0 < \sigma_{\min}^2 < \sigma_{\max}^2 < \infty$ such that

$$\sigma_{\min}^2 \leq \sigma^2(x) \leq \sigma_{\max}^2 \quad \text{for all } x \in \mathcal{X}.$$

- **Higher-order moment bound.** There exists a constant $C < \infty$ and an exponent $\delta > 0$ such that

$$\sup_{x \in \mathcal{X}} \mathbb{E}[|\epsilon_i|^{2+\delta} \mid X_i = x] \leq C.$$

The first condition ensures that matched units have similar variances. Specifically, for any matching scheme with $\|X_{tj} - X_t\| \rightarrow 0$ (as guaranteed by Assumption 4), we have $\sigma^2(X_{tj}) \rightarrow \sigma^2(X_t)$. Hence, $\sigma_j^2 \approx \sigma_t^2$ for $j \in \mathcal{C}_t$ whenever $\mathcal{C}_t$ is constructed by matching on X . In particular,

$$\max_{j \in \mathcal{C}_t} |\sigma^2(X_{tj}) - \sigma^2(X_t)| \rightarrow 0,$$

provided that $\max_{j \in \mathcal{C}_t} \|X_{tj} - X_t\| \rightarrow 0$. *Remark:* This generalizes Assumption 4.1 in [1], which assumes Lipschitz continuity.

The second condition ensures that the conditional variance is bounded away from both zero and infinity. The lower bound prevents degeneracy in the asymptotic distribution and ensures that confidence intervals have positive width. While it is theoretically possible for the variance to approach zero, this would imply that all outcome variability is explained by covariates and there is no residual noise — i.e., $\sigma^2(x) \rightarrow 0$ for all x . In this case, the contribution of the error term to the sampling variability of $\hat{\tau}$ would vanish. The resulting inference problem becomes degenerate: estimation is still possible, but all uncertainty would stem entirely from treatment effect heterogeneity, not from residual noise. This setting is simpler but often unrealistic, as in practice we typically expect some irreducible measurement error in outcomes. The upper bound limits the influence of outliers, which is needed to establish convergence rates.

The third condition imposes a uniform bound on a higher-order conditional moment

of the errors. This assumption is standard in high-dimensional estimation and facilitates the use of maximal inequalities and uniform convergence tools.

We now formally state the assumption we make on the conditional variances of the potential outcomes:

Assumption 5 (Regular error variances) *We assume that both $\sigma_0^2(x)$ and $\sigma_1^2(x)$ are regular variance functions.*

Together, these three properties ensure that both the level (expected magnitude of the errors) and the variability (how much the errors fluctuate around their means) of the error process are well-behaved across the full range of covariates. This structure plays a key role in enabling consistent variance estimation under matching, as we will see in the following sections.

3.10.3 THE FORM OF THE ASYMPTOTIC VARIANCE

We now analyze the asymptotic variance of the matching estimator $\hat{\tau}$. Recall from Equation (3.13) that the estimation error decomposes as:

$$\hat{\tau} - \tau = B_n + E_n + P_n,$$

where B_n captures bias from covariate mismatch, E_n captures sampling error due to residual outcome noise, and P_n captures the discrepancy between the sample and population average treatment effect on the treated (ATT). As discussed earlier, under appropriate regularity conditions and bias correction, the asymptotic distribution of $\hat{\tau}$ is primarily governed by the variability in the two stochastic terms: E_n and P_n .

We are therefore interested in the asymptotic variance of $\hat{\tau}$ as determined by:

$$\text{Var}(\hat{\tau} - \tau) \approx \frac{V_E}{n_T} + V_P,$$

where V_E reflects the contribution from measurement noise (variance of E_n), V_P reflects the population-level heterogeneity in treatment effects among treated units (variance of P_n).

MEASUREMENT ERROR COMPONENT V_E . We first consider the component due to residual outcome noise. Conditional on the covariates $\mathbf{X}$ and treatment assignment vector $\mathbf{Z}$, the variance of E_n is given by:

$$\begin{aligned} V_E &:= \mathbb{E}[E_n^2 \mid \mathbf{X}, \mathbf{Z}] \\ &= \frac{1}{n_T^2} \left(\sum_{t \in \mathcal{T}} \sigma_{1,t}^2 + \sum_{j \in \mathcal{C}} (w_j)^2 \sigma_{0,j}^2 \right), \end{aligned} \tag{3.15}$$

where $w_j = \sum_{t \in \mathcal{T}} w_{jt}$ is the total weight assigned to control unit j across all matched treated units.

This decomposition reflects how residual variance enters the estimator: treated units contribute through their own variances, and controls contribute via squared weight accumulation. Reused controls (with large w_j) disproportionately affect the overall variance. Prior work including [59] and [26] use this variance structure to study the bias-variance tradeoff in matching estimators—highlighting that tighter matches (which reduce bias) can increase variance due to heavy reuse of control units.

Importantly, in the central limit theorem (CLT), the contribution of V_E is scaled by n_T . Since $E_n = O_p(n_T^{-1/2})$, we have:

$$\sqrt{n_T}(\hat{\tau} - \tau) \rightsquigarrow \mathcal{N}(0, V_E + V_P),$$

as formally stated in Theorem 3.10.1.

POPULATION HETEROGENEITY COMPONENT V_P . The second term $P_n = \tau_{\text{SATT}} - \tau$ captures how the realized sample of treated units may differ from the target population

of treated units. That is, even if outcomes were observed without error, the sample ATT may deviate from the population ATT due to treatment effect heterogeneity.

To clarify this, define:

$$V'_P = \text{Var}(\tau_{\text{SATT}} \mid \mathbf{Z}) = \frac{1}{n_T} \mathbb{E} [(\tau(X_i) - \tau)^2 \mid Z_i = 1],$$

so that

$$V_P = n_T \cdot V'_P$$

captures the contribution of this sampling variation to the CLT variance.

Putting both pieces together, the asymptotic variance of $\hat{\tau}$ is given by:

$$\text{Var}(\hat{\tau}) \approx \frac{V_E}{n_T} + V_P,$$

which corresponds directly to the decomposition into E_n and P_n discussed above. The variance term V_E arises from residual outcome variation, while V_P reflects how treatment effect heterogeneity among the treated translates into sampling variability. In the next section, we show how to consistently estimate each component from observed data.

3.10.4 THE CENTRAL LIMIT THEOREM

We now present our main asymptotic normality result, which forms the basis for valid inference.

Theorem 3.10.1 (Central Limit Theorem) *Under Assumptions 1, 2, 3, 4 and 5, as $n_T \rightarrow \infty$:*

$$\frac{\sqrt{n_T} (\hat{\tau} - B_n - \tau)}{V^{-1/2}} \xrightarrow{d} N(0, 1),$$

where

$$V = V_E \cdot n_T + V_P.$$

In the special case where the dimension of the covariate space satisfies $k \leq 2$, the bias term B_n becomes negligible at a faster rate, yielding:

$$\frac{\sqrt{n_T}(\hat{\tau} - \tau)}{V^{-1/2}} \xrightarrow{d} N(0, 1).$$

This theorem generalizes the seminal results of [1], which were limited to M-NN matching with uniform weights ($w_{jt} = 1/M$). Our framework makes two significant advances: (a) it accommodates arbitrary matching procedures including radius matching, caliper matching, and optimal matching; and (b) it allows for flexible weighting schemes such as kernel weights, bias-corrected weights, and synthetic control weights. This increased flexibility enables practitioners to choose matching methods that better balance bias reduction and variance minimization for their specific applications.

For practical implementation of inference procedures, we develop a consistent estimator $\hat{V}$ for the asymptotic variance V in the subsequent section. By Slutsky's theorem, this will yield:

$$\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{\hat{V}^{-1/2}} \xrightarrow{d} N(0, 1).$$

This result provides the foundation for constructing asymptotically valid confidence intervals for the treatment effect. The next section will focus on the consistent estimation of the variance components required for implementation.

3.11 THE STANDARD ERROR ESTIMATOR

To establish valid inference for matching estimators, we analyze a standard error estimation strategy that accommodates both homogeneous and heterogeneous error structures. This approach, previously used in establishing local radius matching in [26] lacks thorough theoretical justification. It is different from existing methods by relaxing traditional assumptions while maintaining consistency under general matching procedures.

Our theoretical analysis provides the missing foundation for this estimator’s widespread application.

The organization of this section is as follows: In Section 3.11.1, we introduce the derivative control condition that generalizes previous assumptions; in Section 3.11.2, we present the formal variance estimator and establish its consistency; and in Section 3.11.3, we compare our approach with previous estimators in the literature, highlighting its practical advantages.

3.11.1 DERIVATIVE CONTROL CONDITION

Before introducing the variance estimator, we first establish a key theoretical condition that enables our analysis. The derivative control condition presented below is more general than the Lipschitz continuity assumption on f used in [1], and allows for broader applicability in settings where f' is not uniformly bounded.

Assumption 6 (Derivative control) *Let f be differentiable on the support of X , and denote its derivative by $f' : \mathcal{X} \rightarrow \mathbb{R}$. There exists a constant $M < \infty$ (possibly depending on n) such that, for all $t \in \mathcal{T}$,*

$$\sup_{x \in \mathcal{C}_t} |f'(x)| \cdot r(\mathcal{C}_t) \leq M,$$

or equivalently,

$$\sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f'(x)| \cdot r(\mathcal{C}_t) \right] < \infty.$$

This condition ensures that in regions where the derivative $f'(x)$ is large, the matching clusters $\mathcal{C}_t$ are sufficiently tight—so that the product of local slope magnitude and cluster size remains uniformly bounded. In contrast to the Lipschitz condition, which imposes a global bound on f' , this condition accommodates functions with steep regions, as long as tighter matches are used locally.

To illustrate the practical advantage, consider $f(x) = x^2$ on $[0, 100]$, where $|f'(x)| = 2|x|$ grows with x , and a global Lipschitz constant would be $L = 200$. Such a large constant makes inference difficult in finite samples. Our condition instead allows for looser matches in flatter regions and tighter matches in steeper regions—offering better practical guidance for match design.

Together with the shrinking cluster condition in Assumption 4, this implies:

$$\sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f'(x)| \cdot r(\mathcal{C}_t) \right] \xrightarrow{n \rightarrow \infty} 0,$$

under the mild requirement that f' is continuous and the support of X remains in a compact region. Since f' is fixed (i.e., does not grow with n), and $r(\mathcal{C}_t) \rightarrow 0$ uniformly in t by Assumption 4, this convergence follows directly. This vanishing bound ensures that local linear approximations to f within each matched cluster incur asymptotically negligible error.

With this theoretical foundation established, we now turn to the variance estimator itself and analyze its consistency properties.

3.11.2 PROPOSED VARIANCE ESTIMATOR

In this section, we introduce our standard error estimator for matching estimators. We begin by presenting the underlying modeling assumptions, then develop the formula for our proposed estimator, and finally establish the consistency results.

Assumption 7 (Homoskedasticity and Regular Variance) *We assume that each unit has the same conditional variance under both treatment and control:*

$$\sigma_{0,i}^2 = \sigma_{1,i}^2 = \sigma_i^2.$$

Furthermore, we assume σ_i^2 is regular in the sense of Definition 3.10.2.

This homoskedasticity assumption simplifies the derivation and allows for tractable plug-in variance formulas. While the assumption may seem restrictive—since in practice the variance may differ across potential outcomes—it serves as a useful approximation, especially when matching quality is high and clusters are tight. By assuming a single variance function $\sigma^2(x)$ governs both outcomes, we avoid needing to estimate two separate variance surfaces.

The variance V consists of two components: the measurement error variance V_E and the population heterogeneity variance V_P . We begin by developing an estimator for V_E , which presents more technical challenges, before extending our approach to estimate the full variance V . Our methodology for V_E establishes key techniques that will later be applied to the full variance estimator. In both cases, our primary theoretical contribution is proving consistency of these estimators under general conditions.

A CONSISTENT ESTIMATOR FOR V_E

To build intuition for our approach, let us first consider the special case where the variance function is constant across all covariate values, i.e., $\sigma^2(x) \equiv \sigma^2$. Under this homoskedasticity assumption, the measurement error variance simplifies to:

$$\begin{aligned} V_E &= \frac{1}{n_T^2} \left(\sum_{t \in \mathcal{T}} \sigma^2 + \sum_{j \in \mathcal{C}} (w_j)^2 \sigma^2 \right) \\ &= \sigma^2 \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right), \end{aligned} \tag{3.16}$$

where $\text{ESS}(\mathcal{C})$ is the effective sample size of the weighted control sample:

$$\text{ESS}(\mathcal{C}) = \frac{(\sum_{i \in \mathcal{C}} w_i)^2}{\sum_{i \in \mathcal{C}} w_i^2}. \tag{3.17}$$

This metric quantifies the number of independent observations that would provide equivalent precision under equal weighting [97], and reflects efficiency loss from reusing controls

with varying weights.

Based on this formula, our proposed plug-in estimator for V_E is:

$$\hat{V}_E = S^2 \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right), \quad (3.18)$$

where S^2 is a pooled variance estimator for σ^2 defined across non-singleton matched clusters. Specifically:

$$S^2 = \frac{1}{N_C} \sum_{t \in \mathcal{T}_+} |\mathcal{C}_t| s_t^2 \quad \text{with} \quad N_C = \sum_{t \in \mathcal{T}_+} |\mathcal{C}_t|, \quad (3.19)$$

where $\mathcal{T}_+ = \{t \in \mathcal{T} : |\mathcal{C}_t| > 1\}$ excludes singleton clusters.

For each cluster, the residual variance is computed as:

$$s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} (Y_j - \bar{Y}_t)^2, \quad \text{where} \quad \bar{Y}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} Y_j. \quad (3.20)$$

We now establish that this estimator is consistent.

Lemma 3.11.1 (Consistency of the Pooled Variance Estimator) *Let $\{\mathcal{C}_t, t \in \mathcal{T}\}$ be a collection of matched control sets. Assume Assumptions 4 (Shrinking Clusters), 5 (Regular Variance), and 6 (Derivative Control). Then, as $n_T \rightarrow \infty$:*

$$\left| S^2 - \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right| \xrightarrow{a.s.} 0. \quad (3.21)$$

Proof: See Appendix C.2.

This result shows that even if variances are not constant across units, the pooled estimator S^2 consistently estimates the average variance across treated units. This aligns with the spirit of heteroskedasticity-robust variance estimation in [125], where consistency is achieved via aggregation even under variance heterogeneity.

Lemma 3.11.2 (Asymptotic Equivalence to Error Variance) *Under the same assumptions as Lemma 3.11.1, define:*

$$\hat{V}_{E,\lim} := \left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right) \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right).$$

Then:

$$\left| \hat{V}_{E,\lim} - V_E \right| \xrightarrow{p} 0 \quad \text{as } n_T \rightarrow \infty.$$

Proof: See Appendix C.3.

Although Equation (3.18) was motivated under homoskedasticity, Lemma 3.11.2 shows that the same formula consistently estimates V_E even when variances are heterogeneous. This is because local variance estimates from matched clusters are close to the true $\sigma^2(X_t)$ due to shrinking clusters and regularity of the variance function. Aggregating over many clusters smooths out local errors—mirroring the robustness of White’s heteroskedasticity-consistent variance estimator.

Theorem 3.11.3 (Consistency of the Variance Estimator) *Under Assumptions 4, 5, and 6, the proposed estimator $\hat{V}_E$ is consistent:*

$$\left| \hat{V}_E - V_E \right| \xrightarrow{p} 0 \quad \text{as } n_T \rightarrow \infty.$$

From Lemma 3.11.1, we have $\left| S^2 - \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right| \xrightarrow{a.s.} 0$. Substituting into our estimator formula:

$$\begin{aligned} \hat{V}_E &= S^2 \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right) \\ &= \left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 + o_p(1) \right) \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right) \\ &= \hat{V}_{E,\lim} + o_p(1) \end{aligned}$$

By Lemma 3.11.2, we have $\left| \hat{V}_{E,\text{lim}} - V_E \right| \xrightarrow{p} 0$. Therefore:

$$\begin{aligned} \left| \hat{V}_E - V_E \right| &= \left| \hat{V}_{E,\text{lim}} + o_p(1) - V_E \right| \\ &\leq \left| \hat{V}_{E,\text{lim}} - V_E \right| + |o_p(1)| \\ &\xrightarrow{p} 0 \end{aligned}$$

This theorem provides the key theoretical guarantee of our method: a plug-in variance estimator motivated by homoskedasticity remains consistent even under general heteroskedastic error structures, as long as the regularity conditions are met. Our estimator offers practical advantages in high-overlap settings, where reuse of control units inflates variance—an effect directly captured by the ESS term.

Our non-parametric approach differs from Theorem 1 of [125], which uses a regression-based (semi-parametric) method. While our matching procedure is governed by hyperparameters such as the number of neighbors or the maximum allowed radius, these parameters are not estimated from the data. Consequently, we require Assumption 5 (Regular Variance), especially the continuity condition in Definition 3.10.2, whereas [125] does not need such an assumption. While both proofs share the same overall strategy, but the specific technical details differ: [125]’s argument relies on compactness of the parameter space to bound the difference between the estimator and the truth, whereas we rely on Assumptions 4 (Shrinking Clusters) and 6 (Derivative Control). Further details on this comparison can be found in Appendix C.6.

A CONSISTENT ESTIMATOR FOR V

Building on our analysis of the measurement error variance component V_E , we now develop a consistent estimator for the total variance V . While V_E captures the variance due to residual outcome noise, the complete variance V must also account for treatment effect

heterogeneity among the treated units.

We start by exploring the relationship between the squared deviations of individual treatment effects and the components of the total variance:

$$\begin{aligned}
& E \left[\left(Y_t(1) - \hat{Y}_t(0) - \tau \right)^2 \right] \\
& \approx E \left[(\tau(x) - \tau)^2 \right] + E \left[\varepsilon_t^2 + \sum_{j \in C_t} w_{jt}^2 \varepsilon_j^2 \right] \\
& \approx V_P + \frac{1}{n_T} \left[\sum_{t \in \mathcal{T}} \sigma_t^2 + \sum_{j \in \mathcal{C}} \left(\sum_{t' \in \mathcal{T}} w_{jt'}^2 \right) \sigma_j^2 \right]
\end{aligned}$$

This expectation can also be approximated empirically as:

$$E \left[\left(Y_t(1) - \hat{Y}_t(0) - \tau \right)^2 \right] \approx \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{Y}_t(0) - \hat{\tau} \right)^2$$

By equating these expressions and rearranging terms, we can derive an estimator for V_P :

$$\begin{aligned}
\hat{V}_P & \approx \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{Y}_t(0) - \hat{\tau} \right)^2 \\
& - \frac{1}{n_T} \left[\sum_{t \in \mathcal{T}} \hat{\sigma}_t^2 + \sum_{j \in \mathcal{C}} \left(\sum_{t' \in \mathcal{T}} w_{jt'}^2 \right) \hat{\sigma}_j^2 \right]
\end{aligned}$$

Combining this with our estimator for V_E , we obtain:

$$\begin{aligned}
\hat{V} & = n_T \hat{V}_E + \hat{V}_P \\
& = \frac{1}{n_T} \left[\sum_{t \in \mathcal{T}} \hat{\sigma}_t^2 + \sum_{j \in \mathcal{C}} \left(\sum_{t' \in \mathcal{T}} w_{jt'} \right)^2 \hat{\sigma}_j^2 \right] \\
& + \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{Y}_t(0) - \hat{\tau} \right)^2 \\
& - \frac{1}{n_T} \left[\sum_{t \in \mathcal{T}} \hat{\sigma}_t^2 + \sum_{j \in \mathcal{C}} \left(\sum_{t' \in \mathcal{T}} w_{jt'}^2 \right) \hat{\sigma}_j^2 \right]
\end{aligned}$$

Through algebraic simplification, this expression reduces to:

$$\begin{aligned}\hat{V} = & \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{Y}_t(0) - \hat{\tau} \right)^2 \\ & + \hat{\sigma}^2 \frac{1}{n_T} \left[\sum_{j \in \mathcal{C}} \left[\left(\sum_{t' \in \mathcal{T}} w_{jt'} \right)^2 - \left(\sum_{t' \in \mathcal{T}} w_{jt'}^2 \right) \right] \right]\end{aligned}$$

where $\hat{\sigma}^2$ is the pooled variance defined in Equation 3.19. This estimator effectively combines the empirical squared deviations with a correction term that accounts for the matching structure.

Theorem 3.11.4 (Consistency of the Total Variance Estimator) *Under Assumptions 4, 5, and 6, the proposed estimator $\hat{V}$ is consistent:*

$$\left| \hat{V} - V \right| \xrightarrow{p} 0 \quad \text{as } n_T \rightarrow \infty.$$

The proof follows similar steps to those used in establishing the consistency of $\hat{V}_E$ in Theorem 3.11.3. The key insight is that both the empirical squared deviations and the correction term converge to their respective population counterparts in probability, leveraging the properties of shrinking clusters, regular error variance functions, and our derivative control condition.

This consistency result ensures that confidence intervals constructed using $\hat{V}$ will have asymptotically correct coverage, providing practitioners with reliable inference tools for matching estimators across a wide range of applications.

3.11.3 COMPARISON WITH [1]ESTIMATOR

To position our work within the existing literature and highlight its advantages, we now compare our variance estimator with that proposed by [1]. This comparison is particularly relevant as their work established the foundational theory for matching estimators, and

our analysis builds upon and extends their approach for practical applications in modern causal inference settings.

Adapting their estimator to our notation:

$$\widehat{V}_{AI06} = \frac{1}{n_T^2} \sum_{t \in \mathcal{T}} \hat{\sigma}_t^2 + \frac{1}{n_T^2} \sum_{j \in \mathcal{C}} \left(\sum_{t \in \mathcal{T}} w_{jt} \right)^2 \hat{\sigma}_j^2, \quad (3.22)$$

where $w_{jt} = 1/M$ if unit j is among the M closest controls to unit t , and $w_{jt} = 0$ otherwise, and $\hat{\sigma}_i^2$ is an estimate of the conditional outcome variance for unit i , defined as:

$$\hat{\sigma}_i^2 = \frac{M}{M+1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{m(i)} \right)^2.$$

Here, $Y_{m(i)}$ denotes the outcome of the m -th closest unit to unit i among units with the same treatment status, and M is a fixed small number (typically set to match the number of matches used in the estimator).

The fundamental difference is in variance estimation: while [1] estimates variance by comparing each unit to its nearest same-treatment neighbors individually, our approach calculates variance within entire matched clusters. This cluster-based approach leverages information from all control units matched to a treated unit simultaneously, resulting in more stable variance estimates, especially when match quality is heterogeneous.

To clarify the key methodological differences, let's compare the variance components directly. In Abadie and Imbens' approach, $\hat{\sigma}_i^2 = \frac{M}{M+1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{m(i)} \right)^2$ estimates variance by comparing each unit to its nearest same-treatment neighbors individually. By contrast, our estimator uses $s_t^2 = \frac{1}{|\mathcal{C}_t|-1} \sum_{j \in \mathcal{C}_t} (Y_j - \bar{Y}_t)^2$, which calculates variance within entire matched clusters. Our approach effectively pools information across all controls matched to a treated unit, leveraging their collective variation to estimate outcome variability more robustly, particularly in settings with variable match quality.

This use of all controls in variance estimation provides several key advantages. First,

our estimator requires matching only for treated units, whereas [1] requires matching for both treated and control units—significantly reducing computational burden when the control group is large. Second, [1]’s approach necessitates matching treated units with other treated units, which becomes problematic when the treated group is small or highly heterogeneous, as is common in many applications. Our approach is tailored specifically for ATT estimation in these realistic scenarios. Third, our framework naturally accommodates flexible weighting schemes, including kernel weights, caliper matching weights, and optimal transportation weights, whereas [1]’s approach was primarily designed for fixed-number nearest neighbor matching with equal weights.

While our estimator does not utilize within-treated-group variation for variance estimation (i.e., our approach does not use the observed outcomes Y_t of treated units when estimating variance), this potential efficiency loss is typically minor in ATT applications where the treated group is small relative to the control group. Furthermore, within-treated-group variation becomes unreliable when the number of treated units is small, making our approach more robust in such common scenarios. The limitation of not using treated outcomes for variance estimation is typically minor in ATT applications where the treated group is small relative to the control group—a common scenario in practice. Indeed, many influential ATT applications feature relatively small treated samples, including job training program evaluations [?], educational interventions [?], and health policy assessments [61]. [57] notes that ATT estimation is often preferred precisely in contexts where treatment is relatively rare or targeted, resulting in small treated groups compared to the potential control pool.

3.12 SIMULATION

In this section, we conduct simulation studies to validate the two main theoretical results established in earlier sections: Theorem 3.10.1 (Central Limit Theorem) and the

consistency of our variance estimator. The primary focus is threefold: first, to verify the asymptotic normality of our estimator, second, to assess whether confidence intervals constructed using our variance estimator achieve near-nominal coverage, thereby confirming the accuracy of inference procedures built from it, and third, to compare the performance of our variance estimator to that of existing methods, showing how our approach outperforms the bootstrap variance estimator proposed by [93]. These simulations provide empirical insights into the reliability and robustness of our methods under different data-generating scenarios and matching conditions.

3.12.1 SIMULATION TO VERIFY THE THEORETICAL RESULTS

In this subsection, we conduct a simulation study to verify the Central Limit Theorem (CLT) result (Theorem 3.10.1) and demonstrate the consistency of our variance estimator, as proper inference depends critically on accurate variance estimation. Following the simulation setup from [26], we generate treated and control units from bivariate distributions with varying degrees of overlap and measure the coverage of the resulting confidence intervals. The complete details of the data-generation process, including specific distribution parameters and matching procedures, are provided in Appendix C.7.1.

We first compute the coverage of a 95% confidence interval using the theoretically correct variance V by constructing $(\hat{\tau} - B_n) \pm \sqrt{\frac{V}{n_T}}$ over 500 replications. The results are presented in Table 3.12.1. The coverage rates for varying degrees of overlap range from 94% to 95.2%, demonstrating that the nominal 95% level is closely attained. We also plot the distribution of $\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{\sqrt{V-1/2}}$ against a standard normal distribution in Figure 3.12.1. We observed a close alignment between these distributions. This provides intuitive evidence that our estimator's distribution converges to the theoretical limit.

We then repeat the coverage analysis using an estimated variance $\hat{V}$ rather than the true variance V . The results remain consistently close to 95% across all overlap levels,

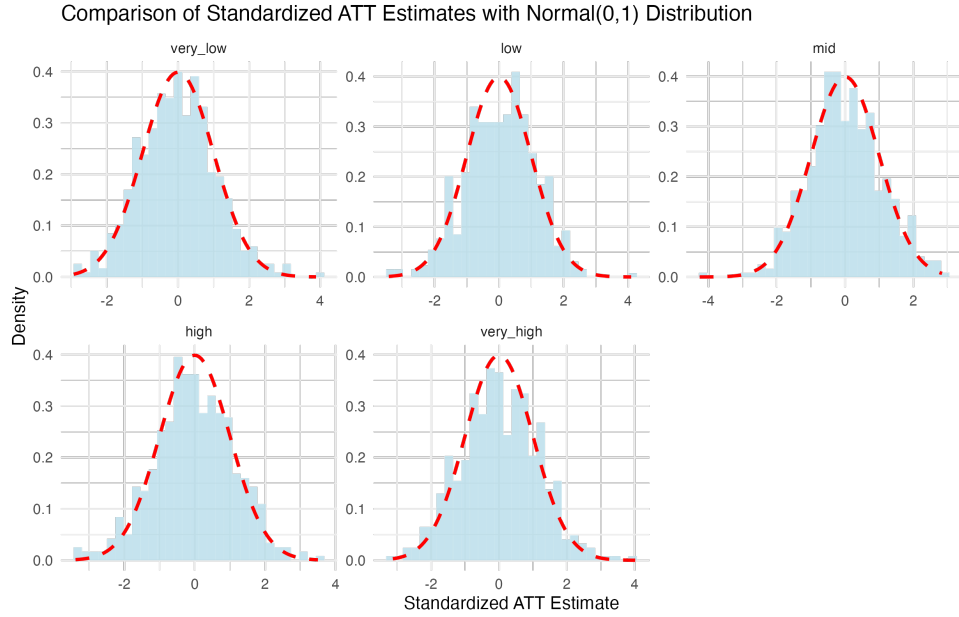


Figure 3.12.1: Empirical distribution of $\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{V^{-1/2}}$ versus the standard normal.

demonstrating that our variance estimation method is accurate even under varying degrees of overlap. This is an indication of robustness of the variance estimator.

Table 3.12.1: Coverage of 95% Confidence Intervals Using True Variance V and Estimated Variance $\hat{V}$

Degree of Overlap	Coverage (%) with V	Coverage (%) with $\hat{V}$
very low	94.0	95.0
low	95.0	94.0
mid	95.2	96.0
high	94.0	95.0
very high	95.0	94.0

These findings confirm that the CLT is well-supported empirically across various overlap scenarios. Moreover, our proposed variance estimator demonstrates reliable coverage properties in practice. The consistently strong performance shown in Table 3.12.1 is particularly noteworthy given the results that follow in the next subsection. In the following comparison with alternative methods, we will demonstrate that while our variance estima-

tor maintains reliable coverage across different degrees of overlap, the bootstrap method struggles substantially in certain scenarios, highlighting the practical superiority of our approach in realistic matching conditions.

3.12.2 A COMPARISON TO THE BOOTSTRAP VARIANCE ESTIMATOR

After validating the CLT, we now compare our variance estimator with the bootstrap variance estimator. To evaluate our proposed methods, we replicate a specific case from [93], focusing on a two-dimensional setting with a complex nonlinear outcome function. The data generating process consists of a treatment assignment mechanism governed by parameters $\gamma_1 = 0.15$ and $\gamma_2 = 0.7$, with treatment probability determined by $P(X_i) = \gamma_1 + \gamma_2 \|X_i\|$. The outcome follows a nonlinear pattern defined by $m(z) = 0.4 + 0.25 \sin(8z - 5) + 0.4 \exp(-16(4z - 2.5)^2)$, where potential outcomes are generated as $Y_i(1) = \tau + m(\|X_i\|) + \epsilon_i$ and $Y_i(0) = m(\|X_i\|) + \epsilon_i$, with $\epsilon_i \sim N(0, 0.2^2)$ and true treatment effect $\tau = 0$. Full simulation setting is provided in Appendix C.7.2.

Our analysis employs 8-nearest neighbor matching with uniform weights of $1/8$ assigned to each matched control unit. As reported in the paper, the true 95% confidence interval (CI) coverage is 0.9473, with an average CI length of 0.2381. Through a comparison of our pooled variance estimator against the wild bootstrap method (see Appendix C.7.2 for details on the wild bootstrap implementation) proposed by [93] across 100 replications, we find differences in performance and present in Table 3.12.2. The wild bootstrap method achieves only 61% coverage with an average confidence interval length of 0.1561, which is substantially shorter than the true interval length of 0.2381. In contrast, our method maintains coverage closer to the nominal rate.

The poor performance of the wild bootstrap can be attributed to the substantial overlap in the matching structure. With an average of approximately 50 units in each of the treatment and control groups, we observe controls being shared across an average of

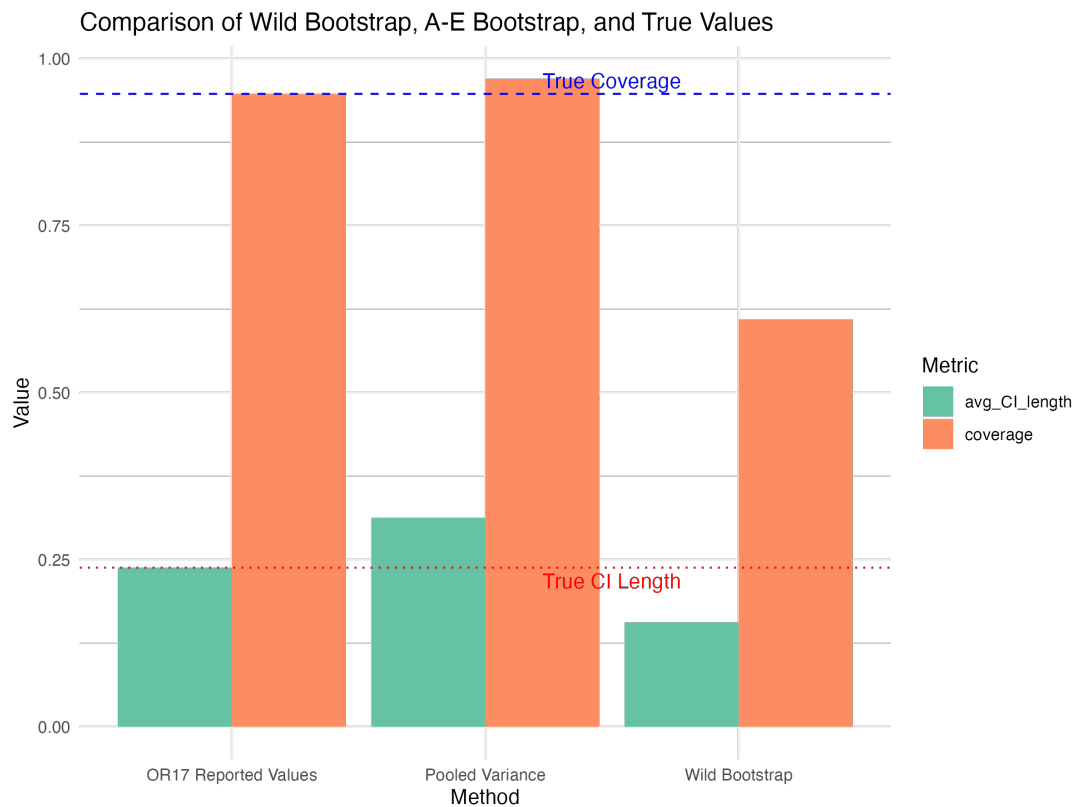


Figure 3.12.2: Comparison of the Wild Bootstrap, A-E Bootstrap, and true values for confidence interval (CI) coverage and average CI length. The Wild Bootstrap method achieves shorter CIs but at the cost of lower coverage, while the A-E Bootstrap method closely approximates the true coverage but results in longer intervals. Horizontal dashed and dotted lines indicate the true coverage and CI length, respectively, for reference.

Method	Coverage (%)	Average CI Length	Difference from True CI Length
Wild Bootstrap	61.00	0.1561	-0.0820
Our Method	97.00	0.3127	+0.0746

Table 3.12.2: Comparison of confidence interval performance across methods. The Wild Bootstrap method has a lower coverage and shorter intervals compared to the true values, while the A-E Bootstrap method achieves coverage closer to the true value but at the cost of longer intervals.

92 treated-control pairs out of approximately 400 total pairs (roughly 25%), indicating extensive reuse of control units in our two-dimensional matching setting. Our analysis shows consistent patterns across different inference methods, with an average of 23 shared controls per treated unit (specifically, 23.17 and 23.45 for wild bootstrap and our method respectively), while the average number of shared treated units per control is approximately 92 (92.63 and 92.23 respectively). This high degree of overlap is not unexpected given that the numbers of treated (n_T) and control (n_C) units are of similar magnitude. To investigate this relationship further, we conducted an additional experiment where we reduced the overlap by setting $N_T = 25$ and $N_C = 1000$. Under these conditions, the wild bootstrap recovers its nominal coverage, while our pooled variance estimator continues to perform well. These findings support our hypothesis that the wild bootstrap procedure is asymptotically valid primarily in settings with minimal overlap.

3.13 APPLICATION OF THE VARIANCE ESTIMATOR TO OTHER ESTIMATORS

While our variance estimator was developed in the context of matching methods, its utility extends to other classes of estimators that share similar properties. In this section, we demonstrate how our approach can be applied to weighting estimators, specifically the stable balancing weights method proposed by [134].

3.13.1 STABLE BALANCING WEIGHTS ESTIMATOR

The stable balancing weights approach of [134] finds weights that minimize the variance of the weighted estimator while satisfying covariate balance constraints. Using our notation, the stable balancing weights estimator for the ATT can be expressed as:

$$\hat{\tau}_{SBW} = \frac{1}{n_T} \sum_{t \in \mathcal{T}} Y_t - \frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j Y_j \quad (3.23)$$

where w_j are weights assigned to control units that minimize $\sum_{j \in \mathcal{C}} w_j^2$ subject to balance constraints of the form $\left| \sum_{t \in \mathcal{T}} \frac{1}{n_T} X_t - \sum_{j \in \mathcal{C}} w_j X_j \right| \leq \delta$ for some small tolerance δ .

The variance estimator for the stable balancing weights approach directly extends our framework. We can apply our estimator from Equation 3.18 with only one modification: while the weights w_j are determined through quadratic optimization rather than matching, the construction of the heteroskedastic variance component S^2 still requires forming local neighborhoods through matching. This hybrid approach leverages the computational advantages of both techniques—optimal weights from the balancing procedure and accurate variance estimation from local matching—resulting in valid inference for the weighting estimator.

3.13.2 SIMULATION STUDY: [60] DGP

To evaluate the performance of our variance estimator when applied to the stable balancing weights estimator, we conducted a simulation study using the data generating process (DGP) proposed by [60]. This DGP is widely used in the causal inference literature as a challenging benchmark due to its non-linear relationships between covariates, treatment, and outcomes.

The Kang and Schafer DGP generates four standard normal covariates (X_1, X_2, X_3, X_4) and then creates non-linear transformations to produce observed covariates. The treat-

ment assignment is a function of these covariates, and the outcome model includes interactions between treatment and covariates, creating a complex setting where many estimators struggle to achieve proper coverage. Full mathematical details of this DGP are provided in Appendix C.7.3.

3.13.3 RESULTS AND DISCUSSION

We applied the stable balancing weights estimator with our variance estimation approach to the Kang and Schafer DGP over 100 independent replications. The 95% confidence intervals constructed using our variance estimator achieved a coverage rate of 98%. This slightly conservative coverage indicates that our variance estimator remains valid even when applied to weighting estimators in challenging settings.

The simulation results demonstrate that our variance estimation framework has broader applicability beyond matching estimators. The slightly higher-than-nominal coverage (98% versus the target 95%) can be attributed to the steep gradient in the Kang and Schafer outcome model, which creates larger effective bias terms that are not fully accounted for in the first-order approximation.

This slight over-coverage suggests directions for future research, particularly in developing refined variance estimators that can better account for steep derivatives in the outcome model. Potential approaches could include higher-order bias corrections or adaptive methods that estimate the curvature of the outcome model.

Nevertheless, the strong performance of our variance estimator when applied to the stable balancing weights approach demonstrates its flexibility and robustness. This extension opens possibilities for creating a unified framework for inference across various classes of weighting and matching estimators in causal inference.

3.14 CONCLUSION

This paper develops new methods for statistical inference in matching estimators, addressing key challenges in both theoretical foundations and practical implementation. We propose a novel variance estimator that remains valid under extensive control unit reuse and is computationally more efficient than existing approaches. Our theoretical framework generalizes the standard Lipschitz continuity assumption, allowing for broader applicability in settings where the derivative grows moderately but not uniformly. Additionally, we resolve the ongoing debate about bootstrap validity for matching estimators by carefully analyzing the conditions under which different bootstrap procedures succeed or fail.

Through extensive simulation studies, we demonstrate that our variance estimator achieves more reliable inference than existing methods do, particularly in settings with substantial overlap between matched samples. Our findings show that the widely used Wild bootstrap approach underperforms in such scenarios, leading to underestimated standard errors and invalid confidence intervals. In contrast, our proposed method maintains accurate coverage even in high-overlap settings, reinforcing its robustness and practical applicability.

Empirical applications suggest that our approach provides a practical and theoretically grounded solution for researchers conducting inference in matched designs, with potential applications in economics, epidemiology, and policy evaluation. Future work will explore extensions to high-dimensional covariate spaces, alternative bias correction techniques, and adaptations for time-varying treatment effects in longitudinal matching designs. We also aim to examine alternative resampling techniques that better account for matched-set dependencies while maintaining computational efficiency.

Our contributions advance the state of the art in matching-based inference by providing a computationally efficient, theoretically justified, and empirically validated approach to variance estimation. These results offer important guidance for practitioners using

matching methods and highlight the need for careful consideration of variance estimation strategies in applied settings.

3.14.1 FUTURE WORK

Our research opens several promising avenues for further investigation. First, while our simulation evidence demonstrates the superiority of our approach’s finite-sample performance to that of existing methods, particularly in settings with extensive control unit reuse, a more rigorous theoretical comparison would enhance our understanding of when and why different inference procedures outperform others. Such theoretical analysis could formally characterize the efficiency gains of our approach relative to alternatives under various data-generating processes and matching configurations.

Second, it is worth exploring improvements to bootstrap procedures specifically designed for matching estimators. For instance, adapting moving block bootstrap procedures like those described in [72] could better handle the complex dependency structure induced by overlapping matches. This approach might preserve the intuitive appeal of bootstrap methods while addressing their current limitations in the matching context. By explicitly modeling the correlation structure among matched units, such modified bootstrap procedures could potentially achieve both valid inference and computational efficiency.



Supplementary Materials for Chapter 1

A.1 THE SEQUENTIAL RISK TIME SAMPLING ALGORITHM PARAMETERS USED IN THE HS V2V3 STUDIES

In this section, we explain how we set tuning parameters for the SeqRTS algorithm. First, we explain the rationale behind setting the upper bound of treatment probabilities to 0.2. Then, we describe the model for $\hat{g}_t$. After both were chosen, $\hat{N}_0$ was tuned using the TUNE algorithm from [78]. Finally, we present how well $\hat{g}_t$ performed using data from HSV2V3.

A.1.1 DETERMINE THE UPPER BOUND OF THE TREATMENT PROBABILITIES

In this section, we explain the rationale behind setting the upper bound of treatment probabilities to 0.2. Figure A.1.1 illustrates the fraction of time individuals remain sedentary during different hours of the day. For instance, at 10 AM, the fraction of sedentary time is calculated from 10 AM to 9 PM. This calculation is repeated hourly from 9 AM to 9 PM (i.e., at 9 AM, 10 AM, ..., 8 PM, 9 PM). The purpose of this computation is to predict sedentary behavior for the remainder of the day, rather than just for the upcoming hour. The plot shows a relatively consistent range, fluctuating between 0.50 at the start of the day and 0.55 towards the end. Consequently, a randomization probability of 0.2 (excluding availability constraints) results in an average of approximately $144 \times 0.525 \times 0.2 = 15.12$ treatments per day. This upper limit is chosen to prevent the risk of under-treatment. The intention is to circumvent scenarios where only a few sedentary periods occur towards the end of the day. In such cases, it is crucial to avoid excessively high probability of intervention. Setting the probability at 0.2 ensures that, even with as few as 10 sedentary periods in a day, the average treatment constraint is likely to be met.

A.1.2 MODEL FOR $\hat{g}_t$

Now we explain our choice of model for $\hat{g}_t$. Given the current run length R_t , define K to be the remaining amount of sedentary “run to go” in the current state (so $R = K + R_t$ is the total run length). We then define F to be the fraction of time in the state “Sedentary” remaining in the day, as shown in Figure A.1.1. Then the amount of remaining time the participant is classified as “Sedentary” is given by $\Omega = K \wedge r + F(r - K)_+$, where $\wedge$ denotes minimum operator and $(x)_+ = \max(x, 0)$ and r denotes the amount of time remaining in the current day. That is, we account for the fact that the expected run length may

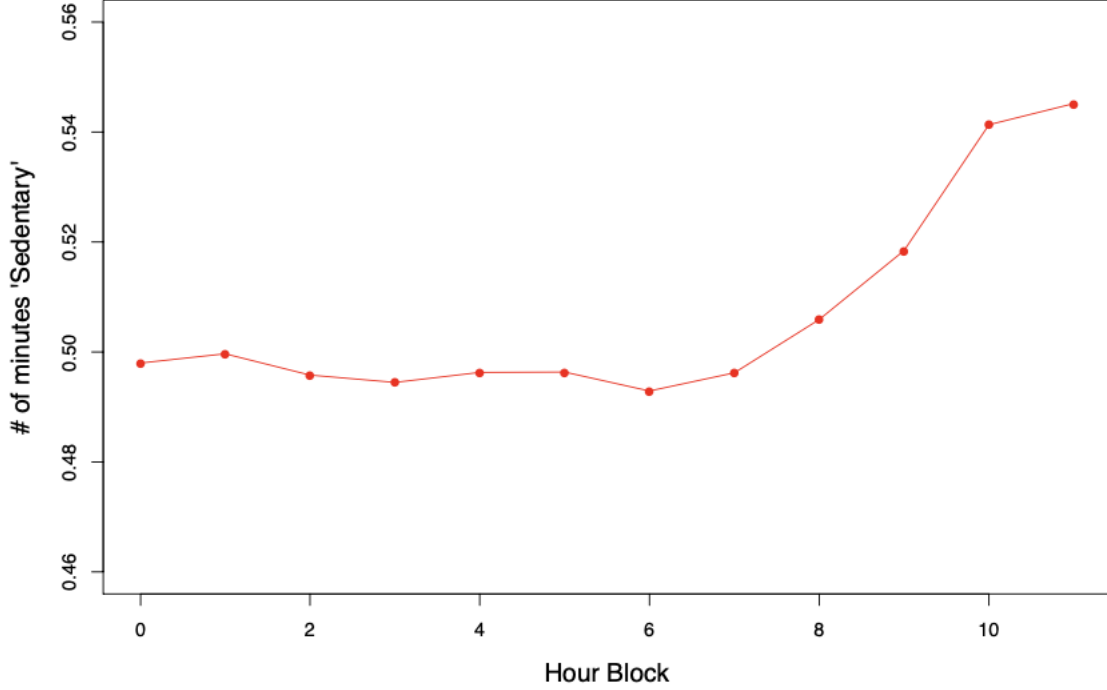


Figure A.1.1: Fraction of time sedentary within and after current hour

be greater than the remainder of the day. We assume that F is independent of both functionals of K given k and B_s . The conditional expectation can then be written as

$$\mathbb{E}[\Omega \mid R_s, B_s] = \mathbb{E}[K \wedge r \mid R_s, B_s] + \mathbb{E}[F \mid R_s, B_s] \cdot \mathbb{E}[(r - K)_+ \mid R_s, B_s]$$

Below we assume $B_s = 1$ as we only intervene when the individual is classified as "Sedentary". Each component of the conditional expectation is estimated using the Heartsteps V1 data. Figure A.1.2 is a Histogram of run lengths in state "Sedentary" and "Not Sedentary" respectively. We checked the Heartsteps V1 data for differences as functions of "time of day" and "length in study", but found very little to suggest a more complicated model was necessary.

We first estimate $\mathbb{E}[K \wedge r \mid R_s = k, B_s = 1]$ and $\mathbb{E}[(r - K)_+ \mid R_s = k, B_s = 1]$. Let $\{\tilde{R}_i\}_{i=1}^m$ denote the data of run lengths used to compute the "Sedentary" plot within Fig-

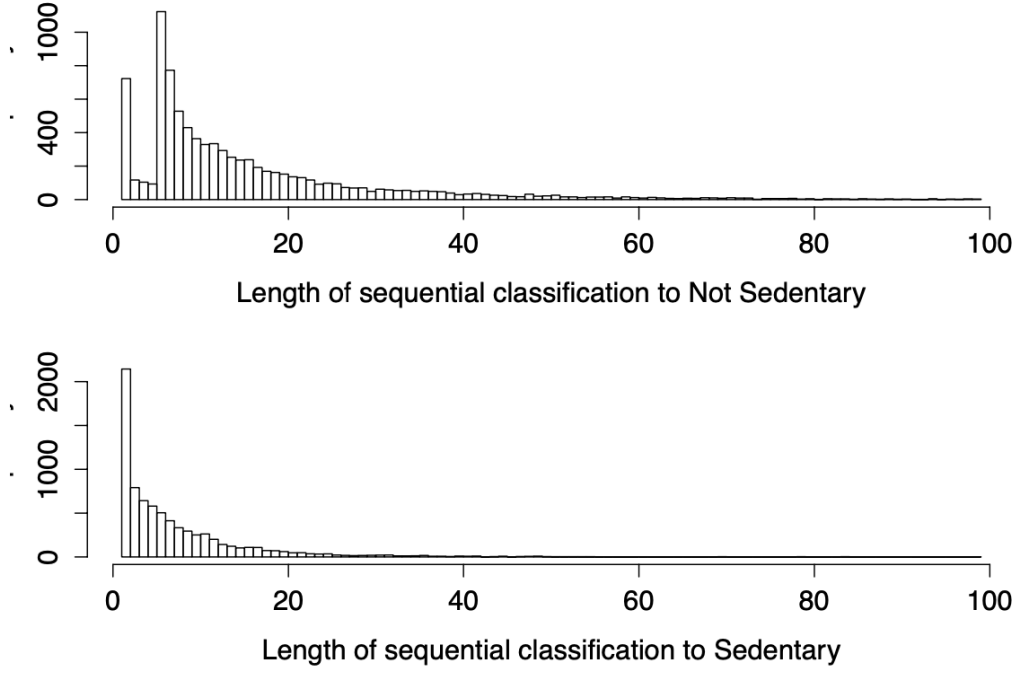


Figure A.1.2: Run lengths of being classified as either “Not Sedentary” or “Sedentary” respectively.

ure A.1.2 as $\tilde{R}_i$. For each $k = 1, 2, \dots, 96$ (two times length of a block), define $\tilde{K}_i = \tilde{R}_i - k$. Using only runs such that $\tilde{K}_i \geq 0$, we have the following estimates

$$\mathbb{E}[K \wedge r \mid k, B_s] \approx \frac{1}{m_k} \sum_{j=1}^{m_k} \tilde{K}_j \wedge r$$

$$\mathbb{E}[(r - K)_+ \mid k, B_s] \approx \frac{1}{m_k} \sum_{j=1}^{m_k} (r - \tilde{K}_j)_+$$

where $m_k \leq m$ is the number of observations satisfying the condition $\tilde{R}_i \geq k$.

We then estimate the fraction of time classified as Sedentary in the remaining window, $\mathbb{E}[F \mid k, B_s]$. Figure A.1.1 plots the estimate calculated using the data from the current hour and all remaining hours. Here, since the figure shows this fraction is flat, we simply use the current hour estimate. For example, at time 9 : 15AM, we use the estimated fraction in Figure A.1.1 from 9 : 00AM—i.e., the fraction of time sedentary from 9 :

00AM to 10 : 00PM. At 9 : 45AM, we use the same estimate; however, at 10 : 00AM and 10 : 10AM, we use the estimated fraction at 10 : 00AM—i.e., the fraction of time sedentary from 10 : 00AM to 10 : 00PM.

A.1.3 QUALITY OF $\hat{g}_t$

This section evaluates the overall effectiveness of $\hat{g}_t$ in predicting the actual remaining risk times g_t . The primary finding is that $\hat{g}_t$ tends to significantly overestimate. We calculated the Root Mean Square Error (RMSE) for each block between $\hat{g}_t$ and g_t . The average RMSEs are presented in Table A.1.1. According to the table, for a median user, $\hat{g}_t$ was overestimated by an average of 3 across all risk times in each block. Considering that a user typically has an average of 7 risk times, this overestimation is substantial.

Table A.1.1: Mean and Mean Absolute difference (MAD) for HSV1 and HSV2V3 — from Liao et al (2018)

	HSV2V3	
	Mean Risk Times	RMSE of $\hat{g}_t$
Block 1	7.79	3.24
Block 2	7.90	3.23
Block 3	8.28	3.17

A.2 A CODING ISSUE THAT AFFECTS EVALUATION

In analyzing HeartSteps V2V3, a coding issue in the algorithm was discovered. This issue makes the SeqRTS algorithm unable to perform as intended. As a result, the algorithm’s performance on both average treatment and uniformity was impacted. We discuss the impact of the coding issue on average treatment and uniformity, the exclusion criteria for data that is largely impacted by this issue, and how it led us to choose data for analysis.

This issue is due to an incorrect conversion of the timezone. There is a need for timezone

standardization because participants have different start times and may live in or travel to different timezones, though most participants lived in Seattle. The code attempts to standardize all start times to a reference time of 8am EST because 8am typically marks the beginning of most participants' day.

Let t^{GMT} be a decision time represented in GMT and t^{EST} be the same time represented in EST. Let t_0^{GMT} represent the starting time in GMT. GMT is involved because it is the default timezone used by R, the coding language that the algorithm employs. The conversion aims to obtain t^{EST} using the formula $t^{EST} = t^{GMT} - (t_0^{GMT} - 14)$, where 14 is chosen because 8am EST corresponds to 2pm (or 14:00) in GMT. This sometimes leads to values outside the expected range when a participant's actual start time deviates from 8am EST, making the $\hat{g}_t$ parameter undefined.

For instance, if a participant begins at 5pm GMT (12pm EST), then at 1am GMT, instead of it being mapped to the correct value 7pm EST, we end up with a calculation of $1am - (17pm - 14pm) = -2$. This makes $\hat{g}_t$ a NA value in the code. This issue occurs in a small portion of block 2 and a significant portion of block 3. All probabilities at these times are set to the algorithm's lower bound of 0.005.

We use the variable `invalid_param_proportion`, mathematically denoted as IPP , as the proportion of risk times when the SeqRTS's parameter becomes invalid due to the coding issue, i.e., the proportion equals to the number of "impacted risk times" divided by 48. Let $k > 0$ be a given threshold value of the proportion. Blocks can be categorized into either $IPP < k$ or $IPP \geq k$. The proportion of the latter category blocks among all blocks is denoted by `Proportion_data_GE.k` (proportion of data greater or equal to k). Analyzing the impact involves constructing confidence intervals for the average treatment and uniformity between the two data categories. We vary k from 0.1 to 0.9. For each fixed k , we present point estimates and confidence intervals of the differences in average treatment and and uniformity separately in Table A.2.1. We also show the proportion of

k	Proportion_data_GE_k	Difference in No. Treatments	Difference in Uniformity
0.1	0.157	-0.109 [-0.136, -0.081]	-0.305 [-0.313, -0.297]
0.3	0.1	0.046 [0.015, 0.077]	-0.336 [-0.347, -0.325]
0.5	0.054	0.310 [0.279, 0.341]	-0.306 [-0.325, -0.286]
0.7	0.03	0.434 [0.403, 0.465]	-0.157 [-0.187, -0.126]
0.9	0.018	0.505 [0.480, 0.531]	0.112 [0.089, 0.135]

Table A.2.1: Comparisons between blocks where $IPP \geq k$ and $IPP < k$. "Proportion_data_GE_k" represents the data proportion with $IPP \geq k$. The point estimate and 95% CI for the difference for these data subsets are presented.

blocks whose $IPP < k$.

The difference between blocks with low IPP and those with high IPP shows that the coding issue indeed influences the algorithm's performance. We choose to exclude entries where $IPP \geq 0.5$ since they introduce a substantial discrepancy between average treatment and uniformity and represent only a small fraction of the dataset.

A.3 CHECKING PLAUSIBILITY OF THE MCAR ASSUMPTION

Note that the evaluation is only conducted on days and blocks when data is not missing for the entire day or block. We acknowledge that the validity of the GEE results relies on the assumption of missing completely at random (MCAR). In this section, we show that this assumption is plausible.

MCAR means the missingness does not depend on any observed variable. However, it might be plausible that some variables can predict the missingness. Two such variables are the weekend and the previous day's number of available times. The weekend might be predictive since it is possible that on weekends, a user has low availability due to not wanting to wear the watch or going on a hike. The previous day's availability might predict the current day's availability because the more that the user is available on the previous day, the more likely it is that the user remembers or wants to wear the watch the next day, i.e., the likelihood they are available is greater.

We use a regression model to check the whether these two variables can predict the missingness. We choose GEE with a logistic link and AR-1 working correlations. The mean model is as follows:

$$\text{logit}P[I_d = 1] = \beta_0 + \beta_1 N_{d-1} + \beta_2 \text{is_weekend}_d \quad (\text{A.1})$$

The results show that β_1 is estimated to be 0.0164 with a p-value less than 10^{-5} , and β_2 is not significant at 0.05 level. This indicates that the more the user is available on the previous day, the more likely it is that the user is available the next day. The weekend variable does not affect availability.

The above says missing at random (MAR) might be more plausible since there is one observed variable that can predict the availability. To account for MAR, we use weighted GEE (WGEE, [103]) to estimate the marginal performance of the algorithm. The result is presented in Table A.3.1.

The weights used in WGEE are left skewed, with the 1st Quartile at 1.1559, the Median at 1.2002, and the 3rd Quartile at 1.2357. These weights indicate minimal variation in the probability of missingness, suggesting that the observed data is nearly uniformly representative of the full dataset. Consequently, the WGEE results are very close to those obtained from the non-weighted GEE. This indicates that the estimates are robust to the assumption of MAR and do not significantly differ from those under the MCAR assumption. Therefore, we conclude that the impact of the potential MAR mechanism is minimal in our context.

Based on this finding, we proceed with the MCAR assumption in the main paper for simplicity and clarity. This approach allows us to use the non-weighted GEE results without a substantial loss of accuracy or validity.

We have identified two kinds of missing days we are not interested in, but there is one kind of missingness we are interested in (when a participant wears the watch, but there is

no connection to the server). In this analysis, we do not differentiate the types of missing days (mainly because we are unable to do so), but we compare MCAR and MAR across all kinds of missing days. If we were able to differentiate the types of missing days, we would analyze a subset of the data.

Table A.3.1: Average Treatment and Uniformity using GEE and WGEE

		width=			
Block / Day	Measure	Avg. Trt.		Uniformity	
		AR-1	AR-1 (WGEE)	AR-1	AR-1 (WGEE)
Whole Day	Mean	1.534 (1.408, 1.659)	1.521 (1.396, 1.647)	0.019 (0.018, 0.020)	0.019 (0.018, 0.020)
	AR-1 Correlation	0.306 (0.032)	0.306 (0.032)	0.163 (0.021)	0.163 (0.021)
Block 1	Mean	0.553 (0.479, 0.625)	0.551 (0.478, 0.625)	0.010 (0.009, 0.011)	0.010 (0.009, 0.011)
	AR-1 Correlation	0.135 (0.019)	0.135 (0.019)	0.055 (0.016)	0.055 (0.016)
Block 2	Mean	0.540 (0.462, 0.618)	0.539 (0.462, 0.617)	0.011 (0.011, 0.012)	0.011 (0.011, 0.012)
	AR-1 Correlation	0.139 (0.026)	0.139 (0.026)	0.016 (0.018)	0.016 (0.018)
Block 3	Mean	0.547 (0.475, 0.618)	0.548 (0.477, 0.620)	0.016 (0.015, 0.018)	0.016 (0.015, 0.018)
	AR-1 Correlation	0.142 (0.021)	0.142 (0.021)	0.176 (0.027)	0.176 (0.027)

1

A.4 CHECKING THE ASSUMPTION THAT THE MEAN IS CONSTANT ACROSS THE DAYS

In this section, we check the assumption that the mean is constant across the days. We justify this by showing no day trend for both the average number of treatments and the average MAD using both exploratory data analyses (EDA) and tests of significance.

We conduct an exploratory analysis by plotting $\bar{Y}_{j,d} = (1/n) \sum_{i=1; d \in \mathcal{D}_i}^{82} Y_{i,j,d}$, $j = 0, 1, 2, 3$. This is the average of number of treatments on day d in block j , averaged

¹Note that the usual non-weighted GEE coefficient estimates are different from the ones in the main paper. This is because the dataset we use here is a subset of the full dataset to ensure the previous day of the analysis day is available (so the weight, one over the probability of being available, can be estimated).

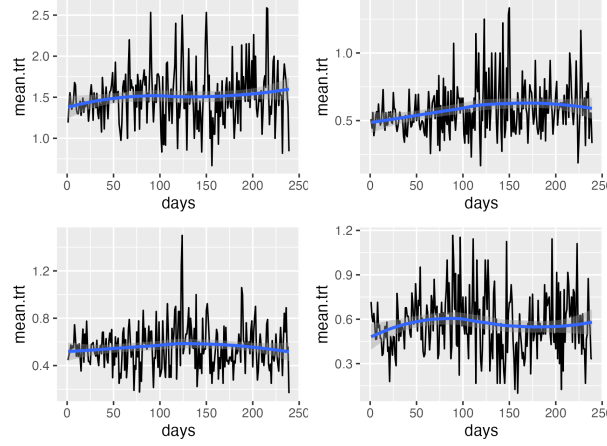


Figure A.4.1: Plot of average of number of treatments over all study days for all blocks. The averages are taken over users who are available on the day.

over users who are available on the day. Given the assumption that data is MCAR (see Appendix A.3), $\bar{Y}_{j,d}$ is a good proxy of the true population average number of treatments on day d . Thus, inspecting $\bar{Y}_{j,d}$ over days d is a way to inspect whether the population mean number of treatment has day trend. On each plot we fix block number j and plot $\bar{Y}_{j,d}$ over days d . We do the same for all blocks separately.

Figure A.4.1 shows the plot. The variation is larger between day 100 to day 150 because the number of available users in those days are low (due to the server issue described in Section 2). The plot shows that in general, there is no variation or mildly increasing trend across days.

To validate the hypothesis, we conducted the regression analysis below. This model is a GEE model specified in Equation 2.5 with the study day term (Day_d) added.

$$\log E[Y_{i,j,d}|X_{i,j,d}] = \log \mu_j + \gamma \text{Day}_d \quad (\text{A.2})$$

$$\text{Var}(Y_{i,j,d}|X_{i,j,d}) = \eta E[Y_{i,j,d}|X_{i,j,d}] \quad (\text{A.3})$$

Here, γ represents the day trend. We perform the analysis for each block $j = 0, 1, 2, 3$. All results show that γ is not significant at 0.05 level, indicating there is no day trend.

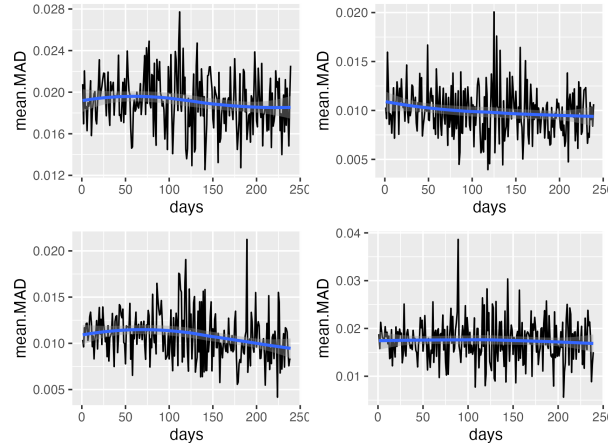


Figure A.4.2: Plot of average MAD over all study days for all blocks. The averages are taken over users who are available on the day.

For MAD, the plots (shown in Figure A.4.2) are generally stable, too. Block 1 shows a slightly decreasing trend. However, when we fit a study day term into the regression, the regression coefficient is not significant, indicating there is no day trend.

In conclusion, the above analysis supports the assumption that the average does not vary over study days for both in all blocks.

A.5 AN ILLUSTRATIVE EXAMPLE OF WHY BLOCK AVERAGES DO NOT SUM TO THE DAILY AVERAGE

The phenomenon where the block goals of 0.5 treatment per block are achieved, but the daily goal of 1.5 treatments per day is not met, occurs because a participant may be available in one block but unavailable in the other two blocks on a given day. Consequently, the number of days counted for the daily average is greater than the number of days counted for the block averages, resulting in a lower daily average compared to the block averages. Figure A.5.1 provides an example illustrating this scenario. For instance, a participant with four study days may be available in all four evening blocks but only available in two morning blocks. When calculating the average treatment in block 1,

	Day 1	Day 2	Day 3	Day 4	Average
Whole Day	0	2	2	1	1.25
Block 1	0	1	-	-	0.5
Block 2	0	1	1	0	0.5
Block 3	0	0	1	1	0.5

Figure A.5.1: This example illustrates why block goals (average of 0.5 treatments per block) are achieved, but the daily goal (average of 1.5 treatments per day) is not. Blue boxes represent blocks or days with at least one available time, and the numbers inside them show the total number of treatments in each block or day. On day 3 and 4, there is no available time in block 1, so the average for block 1 does not include these two days. However, these two days are included in the daily average calculation because there are available times in blocks 2 and 3.

the denominator (the number of available block 1s) is 2. However, when calculating the average treatment per day, the denominator (the number of available days) is 4.

A.6 DISCUSSION ON NOT USING KL DIVERGENCE FOR UNIFORMITY

In past work, [78] proposed evaluating uniformity using Kullback-Leibler (KL) divergence between SeqRTS probabilities and the oracle probabilities (Equation 2.1) at each risk time, then averaging the distance over all risk times in the block. For a day, they use

$$V_d^{KL} = \frac{\sum_{t: X_{d,t}=1} KL(p_{d,t}, p_{d,t}^{Oracle})}{N_d} \quad (\text{A.4})$$

where $KL(p, q) = p \log \left(\frac{p}{q} \right) + (1 - p) \log \left(\frac{1-p}{1-q} \right)$, and $p_t^{Oracle} = 0.5/N_d$.

There are two issues with using KL divergence as a uniformity metric. First, KL divergence is asymmetric, meaning that $KL(p, q)$ and $KL(q, p)$ are generally different. One may use Hellinger distance to address the asymmetry issue. The Hellinger distance, given by $H(p, q) = \frac{1}{\sqrt{2}} \sqrt{(\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2}$, measures how one probabil-

ity distribution Q is different from a second, reference probability distribution P . The corresponding uniformity metric V_d^{Hell} is defined by

$$V_d^{Hell} = \frac{\sum_{t: X_{d,t}=1} H(p_{d,t}, p_{d,t}^{Oracle})}{N_d}. \quad (\text{A.5})$$

Second, uniformity is achieved when V_d^{KL} and V_d^{Hell} are zero, but $V_d^{KL} \neq 0$ (respectively, $V_d^{Hell} \neq 0$) even when all treatment probabilities are equal because both KL divergence and Hellinger distance encode both the uniformity and average treatment constraints. We have shown this using an example in Section 4.2 of the main paper. Here we provide a mathematical derivation.

The goal is to show V_d^{Hell} contains the difference in adhering average 0.5 per block, which we denote as $d(p^{\text{SeqRTS}}, p^{\text{Unif}})$. Note that

$$d(p^{\text{SeqRTS}}, p^{\text{Unif}}) = |(\sum_t p_{d,t}^{\text{SeqRTS}} - 0.5) - (\sum_t p_{d,t}^{\text{Unif}} - 0.5)| = |\sum_t (p_{d,t}^{\text{SeqRTS}} - p_{d,t}^{\text{Unif}})|$$

since $\sum_t p_{d,t}^{\text{Unif}} = 0.5$

We show $V_d^{Hell} \geq C d(p^{\text{SeqRTS}}, p^{\text{Unif}})$ where $C > 0$ is a positive constant.

Fix a t where $I_t = B_t = 1$. Denote $p_{d,t}^{\text{SeqRTS}} = p$ and $p_{d,t}^{\text{Unif}} = q$, we have

$$\begin{aligned} H(p, q) &= \frac{1}{\sqrt{2}} \sqrt{(\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2} \\ &\geq \frac{1}{2\sqrt{2}} (|p - q| + |(1-p) - (1-q)|) \text{ by Holder's Inequality} \\ &= \frac{1}{\sqrt{2}} |p - q| \end{aligned}$$

i.e. $\sqrt{2}H(p, q) \geq |p - q|$

Plug in $p_{d,t}^{\text{SeqRTS}} = p$ and $p_{d,t}^{\text{Unif}} = q$, and sum over all time, we have

$$\begin{aligned}
V_d^{\text{Hell}} &= \frac{1}{N_d} \sum_{t=1,2,\dots,T, \mathbf{1}(I_t=1, B_t=1)} H(p_{d,t}^{\text{Unif}}, p_{d,t}^{\text{SeqRTS}}) \\
&\geq \frac{1}{N_d} \sum_{t=1,2,\dots,T, \mathbf{1}(I_t=1, B_t=1)} \frac{1}{\sqrt{2}} |p_{d,t}^{\text{Unif}} - p_{d,t}^{\text{SeqRTS}}| \\
&\geq \frac{1}{N_d} \frac{1}{\sqrt{2}} \left| \sum_{t=1,2,\dots,T, \mathbf{1}(I_t=1, B_t=1)} (p_{d,t}^{\text{SeqRTS}} - p_{d,t}^{\text{Unif}}) \right| \text{ by Triangle Inequality} \\
&= \frac{1}{N_d} \frac{1}{\sqrt{2}} d(p^{\text{SeqRTS}}, p^{\text{Unif}})
\end{aligned}$$

Since $N_d \leq 48$, we can take $C = \frac{1}{48\sqrt{2}}$

A.7 DETAILS ON MANAGING BASELINE VARIABLES

A.7.1 DETAILS ON CATEGORIZING BASELINE SEDENTARINESS AND BASELINE ACTIVE-NESS

The categorization of step counts within the specific 12-hour window of 8 am to 8 pm offers a nuanced perspective into an individual's activity patterns during standard waking hours. Here we discuss how we categorize activity level using the median daily total number of steps during the baseline period and categorize sedentary behavior using the hours a participant typically spends sitting on a weekday in the past 7 days.

The categorizations of activity level are as follows:

- $\leq 5,000$ steps: Suggestion of low activity level. This aligns with findings that associate fewer than 5,000 steps a day with a sedentary lifestyle [117].
- 5,001 to 7,500 steps: A representation of moderate activity level, also suggested by [117].
- 7,501 to 9,000 steps: Suggestion of high activity level. The upper limit of 9,000,

rather than the often-quoted 10,000 in behavioural literature, is chosen with consideration of the dataset, where the 3rd quartile is observed at 8,745 steps. This suggests that a significant proportion of the population becomes more active, yet may not consistently hit the well-publicized 10,000 step mark within these 12 hours.

- More than 9,000 steps: Reflects individuals who exhibit very high activity levels during this period, often going beyond typical norms.

The categorization of sedentary behaviour is as follows:

- ≤ 4 hours: This range represents individuals who manifest low sedentary behavior during weekdays. Limiting sitting to 4 or fewer hours showcases a lifestyle that incorporates a significant amount of movement or standing activities. Such behavior is align with recommendations from health experts that emphasize the importance of regular movement throughout the day [20].
- 4.01 to 8 hours: This category suggests a moderate level of sedentary behavior. It is a notable range because it encompasses the typical workday duration for many office jobs. Individuals in this bracket likely have occupations that necessitate prolonged sitting or are in environments where movement is limited during the bulk of the day.
- More than 8 hours: Individuals in this range exhibit a high sedentary lifestyle. Daily sitting durations exceeding 8 hours can exponentially amplify the risks associated with sedentary lifestyles, including chronic diseases and other health issues [94].

A.7.2 DETAIL OF ADDRESSING MISSINGNESS IN BASELINE COVARIATES

Addressing missingness in the candidate baseline covariates, specifically the **sedentary behavior** values absent for 9 users, is crucial for analysis stability. We employed three distinct methods, each resulting in a separate dataset for further evaluation. The missing

sedentary behavior data was managed through: (i) K-Nearest Neighbor (KNN) imputation which gave rise to the `df_KNN` dataset, (ii) median imputation which leads to the `df_median` dataset, and (iii) a straightforward approach where entries with missing data were omitted, producing the `df_drop` dataset. We did not employ multiple imputation due to its potential drawbacks in this context: One primary concern is its complexity, which might not necessarily yield better results compared to simpler methods given the size of our dataset. Multiple imputation might introduce additional uncertainty and does not necessarily ensure enhanced analysis stability. Evaluating our methodologies on these three datasets sufficiently allows us to ensure the robustness and stability of our analysis.

A.8 ENSURE NO ENDOGENEITY FROM THE TIME-VARYING COVARIATES

A.8.1 IMPUTATION OF PARTIAL RISK INDICATOR

Endogeneity arises when an independent variable in a model is correlated with the error term. It is known that endogeneity can affect GEE estimates because it introduces correlation between the predictors and the error term, leading to biased and inconsistent estimates (Diggle 2002). In our paper, caution should be put on the risk time indicator. Specifically, as explained in Section 2.2, a decision time becomes unavailable if the participant has received any type of notification (including an anti-sedentary message) within the preceding hour. This means the risk time is influenced by the outcome.

The perfect indication of a participant’s sedentary behavior, while wearing the watch, can be achieved by eliminating any unavailability of notifications within one hour and adhering to the other three availability criteria. The partial availability indicator, which we denote as $\tilde{I}_t$, is one when only the first three availability criteria are satisfied. Recall that each availability criterion is defined by the converse of its corresponding unavailability cri-

terion. The partial risk indicator $\tilde{X}_t := \mathbf{1}(B_t = 1, \tilde{I}_t = 1)$ provides a comprehensive insight into a participant’s behavior, without any interference from the algorithm’s decision.

Regrettably, $\tilde{I}_t$ were not maintained during the study, except for a brief period at the beginning. Consequently, we are mostly unable to obtain this ideal indication of a participant’s sedentary behavior. To handle this issue, we devised two different strategies to approximate it and opted for the one that minimized the prediction error during the brief period when the indicators for the other three availability criteria were available. The two methods are:

1. *Estimation Method 1*: Impute $\tilde{I}_t = 1$ for all decision times within one hour following the treatment. This is based on the assumption that there is no treatment effect and participants are likely to be sedentary and available within the treated hour, since they have already been sedentary and available.
2. *Estimation Method 2*: Impute $\tilde{I}_t = B_t$. This is based on the assumption that $B_t = 1$ implies availability on the other three criteria, and $B_t = 0$ means the user is not available (i.e., did not wear the watch, was being active, or the watch lost connection).

The indicators of the other three unavailabilities were available before 2019-11-10 (1,555 participant days). These were used as our test set. We calculate the prediction error, defined as the proportion of correct predictions on the test set. Estimation method 1 gives an error of 0.10, and estimation method 2 gives an error of 0.15. Hence, we proceed with estimation method 1.

A.8.2 EFFECT OF THE TREATMENT ON SEDENTARINESS

In this section, we provide data analysis to support that messages sent on one day do not impact sedentariness on the next day. We assess the causal effect of whether there was

one treatment sent on the previous day $d - 1$ on today's sedentariness N_d (number of risk times on day d).

Denote by Z_d the vector of daily characteristic collected after day $d - 1$ and just before the randomization happened on day d . Z_0 represent the baseline covariates. Recall the Y_d represents the number of anti-sedentary messages sent on day d , I_d represents the availability of day d . The treatment variable we are interested is $A_d = \mathbf{1}(Y_d > 0)$, whether or not there is one treatment sent in the previous day $d - 1$. The outcome of interest is N_d , number of risk times (aka the sedentariness) on day d . Data arrives in a sequence of

$$\{Z_0, A_0 = 0; Z_1, I_1, Y_1, A_1, Z_2, I_2, Y_2, A_2; \dots, Z_{d-1}, I_{d-1}, Y_{d-1}, A_{d-1}; Z_d, I_d, Y_d, A_d, \dots, Z_D, I_D, Y_D\}$$

We denote the information occurred up to day d as $H_d = (\bar{Z}_d, \bar{A}_{d-1}, \bar{I}_d)$. In this analysis we assume participants are independent and identically distributed from an unknown distribution P_0 .

To define treatment effects, we use the potential outcomes framework [101, 106]. For a participant, the potential outcomes are

$$\{N_1, A_1, N_2(a_1), A_2(a_1), \dots, N_{D-1}(\bar{a}_{D-2}), A_{D-1}(\bar{a}_{D-2}), N_D(\bar{a}_{D-1})\}$$

We assess the the marginal causal excursion effect (MCEE, [21]) of A_d on N_d :

$$\begin{aligned} \beta = & E\{E[N_d(\bar{A}_{d-2}, 1)|H_d, I_{d-1} = 1]|I_{d-1} = 1\} - \\ & E\{E[N_d(\bar{A}_{d-2}, 0)|H_d, I_{d-1} = 1]|I_{d-1} = 1\} \end{aligned} \tag{A.6}$$

² The MCEE represents the contrast between at least one push notification ($a_{d-1} = 1$) versus no push notification ($a_{d-1} = 0$) on day $d - 1$, following the observed treatment

²Here by $H_d, I_{d-1} = 1$ we mean $H_d/I_d, I_{d-1} = 1$. We use the former notation for conciseness.

history up to day $d - 2$. The quantity marginalizes on all kind of participants and day.

To represent the CEE in terms of the observed quantities, we make the following identification assumptions

Assumption 1 (Consistency) $N_d(\bar{A}_{d-1}) = N_d$, for all $d = 1, \dots, D$. $A_d(\bar{A}_{d-1}) = A_d$, for all $d = 1, \dots, D - 1$.

Assumption 2 (Positivity) If $P(H_d = h_d, I_{d-1} = 1) > 0$, then $P(A_d = a | H_d = h_d, I_{d-1} = 1) > 0$ for $a \in \{0, 1\}$.

Assumption 3 (Sequential ignorability) For $d = 1, \dots, D$, $N_d(\bar{a}_d) \perp\!\!\!\perp A_d | H_d$

With the assumptions, the CEE is identified as

$$\begin{aligned} \beta = & E\{E[N_d | H_d, I_{d-1} = 1, A_{d-1} = 1] | I_{d-1} = 1\} - \\ & E\{E[N_d | H_d, I_{d-1} = 1, A_{d-1} = 0] | I_{d-1} = 1\} \end{aligned} \quad (\text{A.7})$$

To estimate the causal effect, we need a working model for $E\{E[N_d | H_d, I_{d-1} = 1, A_{d-1} = 0] | I_{d-1} = 1\}$. A working model approximates the true conditional expectation as well as possible using variables that predict the outcome but not the treatment (called control variables), though it does not need to be perfectly specified. A good working model will improve the estimation efficiency of the effect. We choose $\alpha \cdot Z_d$ for the working model, where Z_d is the past 5-day average of the number of sedentary times prior to day d , as we believe Z_d predicts N_d , and the model shows that Z_d is significant.

The effect estimate is 0.231 with a p-value of 0.074, indicating no treatment effect. We conclude that there is no effect of treatment on the next day's sedentary behavior, i.e., participants do not respond to the intervention by becoming less sedentary the next day.

A.9 THE ROBUSTNESS OF THE ALGORITHM'S PERFORMANCE

One factor complicating real-world decision-making is participants' unavailability during a one-hour period after each treatment, as mentioned in Section 2.2. Due to this criterion,

even oracle sampling probabilities (see Equation 2.1) may not fulfill the average treatment constraint. This is because the oracle only knows the risk times but not the actual treatment assignments, which are determined by a Bernoulli draw. Consider a block where a user is sedentary for 5 times. The oracle probability is set to $0.5/5 = 0.1$. If a treatment is sent on the third instance, then the fourth and fifth instances would not be allowed to randomize, i.e., the probability will be forced to be 0 instead of 0.1 if these instances fall within the hour after the treatment. In general, a treatment decreases the opportunities for randomization. As a result, the actual average treatment delivered is less than desired. Thus, the oracle approach, which distributes a budget of 0.5 evenly across all risk times in probability, could result in the actual average treatment being lower than 0.5. This is verified in a simulation below.

Though it affects the oracle approach, this real-world complication was addressed by the study designers during the parameter tuning phase of SeqRTS. One way this was handled is by setting $\hat{N}_0$, the tuning parameter of the average treatment goal, to be much larger than the actual goal of 0.5 (recall in Section 3 that $\hat{N}_0$ was set to 1.8).

The above hypothesizes that the oracle fails to achieve the target of 0.5 treatments per block, but SeqRTS is capable of meeting this goal. We now use a simulation to demonstrate this. In the simulation, we go through all available participant blocks in HSV2V3 (refer to Section 4 for the selection process). It is important to note that the availability indicator is dependent on the SeqRTS algorithm due to the one-hour unavailability following treatment. We eliminated this dependence by using an availability indicator based on other criteria, as detailed in Appendix A.8.1. To ensure a fair comparison, we also simulated the performance of SeqRTS instead of using the actual randomization probabilities and treatments from the study.

The results of the simulation comparing SeqRTS and uniform sampling are presented in Figure A.9.1. Indeed, the oracle yielded an average of fewer than 0.5 treatments per block, while the SeqRTS algorithm achieved close to an average of 0.5 treatments

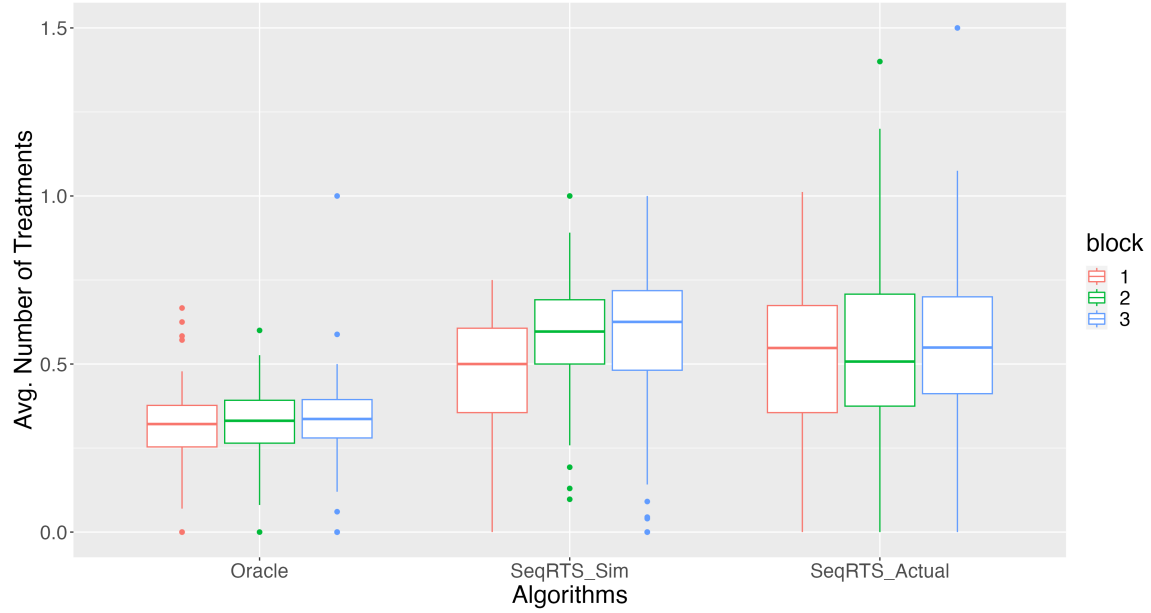


Figure A.9.1: Boxplot of the average number of treatments per participant for each block calculated using SeqRTS and the Oracle. SeqRTS_Sim represents the performance of SeqRTS on the simulated risk times, while SeqRTS_Actual represents the performance of SeqRTS on actual decision times in HSV2V3.

per block. We also presented the average treatment calculated using the actual data in HSV2V3. The performance of the simulated SeqRTS differs from that of the actual one because, in the simulation, there are more risk times (due to the imputation of the risk times in Appendix A.8.1) than actually occurred. Overall, this comparison demonstrates that SeqRTS is robust in real-world environments.

B

Supplementary Materials for Chapter 2

B.1 DERIVATION OF THE TV-CSL ESTIMATOR

The main technique involves expanding the log-likelihood of the data around the true parameter to obtain a score that approximates the true parameter at the fastest possible rate. See the next section for a derivation of the likelihood expansion. A key result from that section is the estimating equation resulting from the optimal score:

$$0 = \mathbb{E} \left[(1(A \leq t) - a_t(X)) X (\Delta - \Lambda(U) e^{\eta_0^*(X) + \tau 1(A \leq t)}) \right] \quad (\text{B.1})$$

Use the tower property to take $E[\cdot | A, X]$ inside, we obtain

$$0 = \mathbb{E} \left[(1(A \leq t) - a_t(X)) X (1 - e^{\eta_A^*(X) - \eta_{a_t}(X)}) \mathbb{P}(\Delta | A, X) \right] \quad (\text{B.2})$$

where

$$\begin{aligned} \mathbb{P}(\Delta | A, X) &= \int_0^A (1 - \exp(-\Lambda(c)e^{\eta_0^*(X)})) f_C(c | A, X) dc \\ &\quad + \int_A^\infty \left(1 - \exp(-\Lambda(A)e^{\eta_0^*(X)} - (\Lambda(c) - \Lambda(A))e^{\eta_0^*(X) + \tau}) \right) f_C(c | A, X) dc \\ &= \int_0^A (1 - \exp(-\Lambda(c)e^{\eta_0^*(X)})) f_C(c | A, X) dc \\ &\quad + \int_A^\infty \left(1 - \exp([- \Lambda(A)(1 - e^\tau) - \Lambda(c)e^\tau]e^{\eta_0^*(X)}) \right) f_C(c | A, X) dc \end{aligned}$$

Note that $\eta_W^*(X) - \eta_W(X) = \nu^*(X) - \nu(X)$. Taking $E[\cdot | X]$ inside, we obtain

$$\begin{aligned} 0 &= \mathbb{E} \left[\left\{ \int_0^t (1 - a_t(X)) \mathbb{P}(\Delta | A = s, X) f(A = s | X) ds + \right. \right. \\ &\quad \left. \int_t^\infty (0 - a_t(X)) \mathbb{P}(\Delta | A = s, X) f(A = s | X) ds \right\} \\ &\quad \left. X \cdot (1 - e^{\nu^*(X) - \nu(X)}) \right] \end{aligned} \quad (\text{B.3})$$

Hence we need

$$\begin{aligned} &(1 - a_t(X)) \int_0^t \mathbb{P}(\Delta = 1 | A = s, X) f(A = s | X) ds \\ &- a_t(X) \int_t^\infty \mathbb{P}(\Delta = 1 | A = s, X) f(A = s | X) ds = 0 \\ &\frac{a_t(X)}{1 - a_t(X)} = \frac{\int_0^t \mathbb{P}(\Delta = 1 | A = s, X) f(A = s | X) ds}{\int_t^\infty \mathbb{P}(\Delta = 1 | A = s, X) f(A = s | X) ds} \end{aligned} \quad (\text{B.4})$$

Here, $\int_0^\infty \mathbb{P}(\Delta = 1 \mid A = s, X) f(A = s|X) ds = P(\Delta = 1 \mid X)$ is the marginal censoring probability.

Hence, $a_t(X) = \frac{\int_0^t \mathbb{P}(\Delta = 1 \mid A = s, X) f(A = s|X) ds}{\int_0^\infty \mathbb{P}(\Delta = 1 \mid A = s, X) f(A = s|X) ds} = \frac{\int_0^t \mathbb{P}(\Delta = 1, A = s|X) ds}{\mathbb{P}(\Delta = 1 \mid X)} = \int_0^t \mathbb{P}(A = s|\Delta = 1, X) ds = \mathbb{P}(A \leq t|\Delta = 1, X)$. When $\mathbb{P}(\Delta = 1 \mid A = s, X) = 1$ for all s , i.e., when all observations are not-censored, then $a_t(X) = \mathbb{P}(A \leq t|X)$ is the treated probability at time t for a unit with covariate X .

B.2 KEY LEMMA FOR DERIVING THE SCORE FUNCTION

Here we derive the key lemma that shows how the estimator is obtained. The key step is to calculate the expansion of the likelihood. Let $\ell(Y; \eta')$ denote the log-likelihood of the exponential family. Lemma 1 of [45] states that for arbitrary η' , the likelihood of Y satisfies

$$\begin{aligned} \ell(Y; \eta') &= \ell(Y; \eta) - \frac{1}{2} \psi''(\eta) (r + \eta - \eta')^2 + \frac{1}{2} \psi''(\eta) r^2 \\ &\quad + O\left(\|\eta - \eta'\|_2^3\right), \end{aligned}$$

where $r := (Y - \psi'(\eta)) / \psi''(\eta)$.

Further

$$\ell(Y; \eta') = \ell(Y; \eta) - \frac{1}{2} \psi''(\eta) (r + \eta - \eta')^2 + \frac{1}{2} \psi''(\eta) r^2 \quad (\text{B.5})$$

The key insight is that we parametrize $\eta_w(x) = \nu(x) + (w - a(x))\tau$, where $\nu(x) = a(x)\tau + \eta_0(x)$. Instead of parametrizing $\eta_w(x)$ using $\eta_0(x)$ and τ , we re-parametrize it by adding and subtracting $a(x)\tau$ to obtain double robustness.

When comparing η and η' , we keep $\nu(x)$ and $a(x)$ fixed, choosing $\eta' = \nu(x) + (w - a(x))\tau'$ and $\eta = \nu(x) + (w - a(x))\tau$. This implies that $\eta' - \eta = (w - a(x))(\tau' - \tau)$. We can then

apply this to Equation B.5:

$$\begin{aligned}
& \ell(Y; \eta') - \ell(Y; \eta) \\
&= -\frac{1}{2}\psi''(\eta)(r + \eta - \eta')^2 + \frac{1}{2}\psi''(\eta)r^2 \\
&= -\frac{1}{2}\psi''(\eta)(r - (w - a(x))(\tau' - \tau))^2 + \frac{1}{2}\psi''(\eta)r^2
\end{aligned} \tag{B.6}$$

Move the negative sign, we have

$$\begin{aligned}
& \ell(Y; \eta) - \ell(Y; \eta') \\
&= \frac{1}{2}\psi''(\eta)(r + (w - a(x))(\tau - \tau'))^2 + \frac{1}{2}\psi''(\eta)r^2
\end{aligned} \tag{B.7}$$

We take the expectation, and maximize the LHS by differentiating τ , and set the derivative to zero (this is to find the tangent direction)

$$0 = E[(w - a(x))\psi''(\eta)(r + (w - a(x))(\tau - \tau'))] \tag{B.8}$$

This says

$$\begin{aligned}
E[(w - a(x))\psi''(\eta)r] &= E[(w - a(x))^2\psi''(\eta)](\tau' - \tau) \\
(\tau' - \tau) &= E[(w - a(x))\psi''(\eta)r] / E[(w - a(x))^2\psi''(\eta)]
\end{aligned} \tag{B.9}$$

We also want $\tau' = \tau$. This is because we are taking one-dimensional efficient scores so we finally need tangency. The condition is that the numerator is zero, i.e.

$$E[(w - a(x))\psi''(\eta)r] = 0 \quad \text{for all } \eta \tag{B.10}$$

Plugin $r := (Y - \psi'(\eta)) / \psi''(\eta)$, and taking x, w as random variables, we have

$$\begin{aligned}
& E[(W - a(X))(Y - \psi'(\eta))] \\
&= E[(W - a(X))(Y - (\psi'(W\eta_1(X) + (1 - W)\eta_0(X))))] \\
&= E\left[(e(X) - a(X))\left(Y - (e(X)\psi'(\eta_1(X)) + (1 - e(X))\psi'(\eta_0(X)))\right)\right]
\end{aligned}$$

B.3 PROOF OF PROPOSITION 1

Proof¹: We write $\gamma_n(x) = \gamma(x) + \alpha_n \xi_n(x)$, where $\mathbb{E}[\xi_n^2(X)] = 1$ is a unit directional vector, and α_n is the distance from $\gamma_n(x)$ to $\gamma(x)$. Similarly, we can write $\eta_n(x) = \eta(x) + \rho_n \zeta_n(x)$, where $\mathbb{E}[\zeta_n^2(X)] = 1$. By the assumption of proposition 1, $\alpha_n \rightarrow 0$, $\rho_n \rightarrow 0$.

The score function for the partial likelihood of the i -th sample is:

$$\begin{aligned}
S_i(\gamma, \eta, \beta) &= S_i(\gamma, \nu, \beta) \quad \text{since } \nu = \eta + \tau \cdot a \\
&:= s(\gamma(X_i), \nu(X_i), \beta) \\
&= \frac{\partial}{\partial \beta} \left[\nu(X_i) + (W_i - a(X_i))X_i^\top \beta \right. \\
&\quad \left. - \log \left(\sum_{l \in \mathcal{R}_i} \exp(\nu(X_l) + (W_l - \hat{a}(X_l))X_l^\top \beta) \right) \right] \\
&= Z_i - \frac{\sum_{l \in \mathcal{R}_i} Z_l \exp(\nu(X_l) + Z_l^\top \beta)}{\sum_{l \in \mathcal{R}_i} \exp(\nu(X_l) + Z_l^\top \beta)}
\end{aligned}$$

where $Z_i := (W_i - a(X_i))X_i$.

Denote the expected score as $s(\gamma, \nu, \beta) = E[S_i(\gamma, \nu, \beta)]$ and define the empirical score $s_n(\gamma, \nu, \beta) = \frac{1}{n} \sum_{i=1}^n S_i(\gamma, \nu, \beta)$. For simplicity, we write $s_n(\gamma_n, \nu_n, \beta_n)$ as $s_n(\alpha_n, \rho_n, \beta_n)$

¹The proof structure follows the approach of [45], which we have independently verified for our specific context of survival analysis with random treatment timing. In subsequent versions, we will extend this foundation with additional martingale arguments that address the survival model present in our setting, further strengthening the theoretical guarantees for our estimator.

We first show β_n is consistent under $s_n(0, 0, \beta)$. Taylor's expansion of $s_n(\alpha_n, \rho_n, \beta_n)$ at $\alpha_n = \rho_n = 0$ is

$$\begin{aligned} & s_n(\alpha_n, \rho_n, \beta_n) \\ &= s_n(0, 0, \beta_n) + \nabla_{\alpha} s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \alpha_n + \nabla_{\rho} s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \rho_n \end{aligned}$$

where $\alpha_{\varepsilon} \in [0, \alpha_n], \rho_{\varepsilon} \in [0, \rho_n]$.

Note that $s(0, 0, \beta_0) = 0$ (See [43], Chapter 8 for a proof). Thus $s_n(0, 0, \beta_n) = s_n(0, 0, \beta_n) - 0 = s_n(0, 0, \beta_n) - s(0, 0, \beta_0)$. We now argue $s_n(0, 0, \beta_n) - s(0, 0, \beta_0) = \nabla_{\beta} s_n(0, 0, \beta_{\varepsilon})(\beta_n - \beta_0) + s_n(0, 0, \beta_0)$ where $\beta_{\varepsilon} \in [\beta_n, \beta]$.

We make the following decomposition:

$$\begin{aligned} & s_n(0, 0, \beta_n) - s(0, 0, \beta_0) \\ &= (s_n(0, 0, \beta_0) - s(0, 0, \beta_0)) + (s(0, 0, \beta_n) - s(0, 0, \beta_0)) + \\ & \quad [(s_n(0, 0, \beta_0) - s(0, 0, \beta_0)) - (s_n(0, 0, \beta_n) - s(0, 0, \beta_n))] \end{aligned}$$

The last term in the bracket is an empirical process term. Given that our score function $s(0, 0, \beta)$ is a Donsker class, and β_n is consistent, the empirical process term is $o_P(n^{-1/2})$ (Lemma 19.24 of [118]).

Furthermore, by mean value theorem,

$$s(0, 0, \beta_n) = s(0, 0, \beta_0) + \nabla_{\beta} s(0, 0, \beta_{\varepsilon})(\beta_n - \beta_0)$$

for some $\beta_{\varepsilon} \in [\beta_n, \beta]$.

Thus,

$$\begin{aligned}
s_n(0, 0, \beta_n) &= s_n(0, 0, \beta_n) - s(0, 0, \beta_0) \\
&= \nabla_{\beta} s_n(0, 0, \beta_{\varepsilon})(\beta_n - \beta_0) + s_n(0, 0, \beta_0)
\end{aligned} \tag{B.11}$$

Furthermore, by central limit theorem (CLT),

$$s_n(0, 0, \beta_0) = s(0, 0, \beta_0) + O_p(n^{-1/2}) = O_p(n^{-1/2})$$

Notice that $\nabla_{\alpha} s_n(\alpha_n, \rho_n, \beta_n)$, $\nabla_{\rho} s_n(\alpha_n, \rho_n, \beta_n)$ are bounded, i.e., they are both $O_P(1)$

$$\begin{aligned}
0 &= s_n(\alpha_n, \rho_n, \beta_n) - s_n(0, 0, \beta_n) = s_n(0, 0, \beta_n) + \\
&\quad \nabla_{\alpha} s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \alpha_n + \nabla_{\rho} s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \rho_n \\
&= \nabla_{\beta} s_n(0, 0, \beta_{\varepsilon})(\beta_n - \beta_0) + O_p(n^{-1/2} + \alpha_n + \rho_n)
\end{aligned}$$

Then, because the minimum eigenvalue of $\nabla_{\beta} s(0, 0, \beta_{\varepsilon})$ and is lower bounded, the above turns to :

$$\beta_n - \beta_0 = (\nabla_{\beta} s_n(0, 0, \beta_{\varepsilon}))^{-1} O_p(n^{-1/2} + \alpha_n + \rho_n) = o_p(1)$$

Therefore, β_n is consistent.

We now prove the rate result. To do this, we make a second order Taylor's expansion of $s_n(\alpha_n, \rho_n, \beta_n)$ at $\alpha_n = \rho_n = 0$:

$$\begin{aligned}
& s_n(\alpha_n, \rho_n, \beta_n) \\
&= s_n(0, 0, \beta_n) + \nabla_{\alpha} s_n(0, 0, \beta_n) \alpha_n + \nabla_{\rho} s_n(0, 0, \beta_n) \rho_n \\
&\quad + \frac{1}{2} \nabla_{\alpha}^2 s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \alpha_n^2 + \frac{1}{2} \nabla_{\rho}^2 s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \rho_n^2 \\
&\quad + \nabla_{\alpha\rho} s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \alpha_n \rho_n \\
&= s_n(0, 0, \beta_0) + \nabla_{\beta} s_n(0, 0, \beta_{\varepsilon}) (\beta_n - \beta_0) + \nabla_{\alpha} s_n(0, 0, \beta_n) \alpha_n \\
&\quad + \nabla_{\rho} s_n(0, 0, \beta_n) \rho_n + \frac{1}{2} \nabla_{\alpha}^2 s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \alpha_n^2 \\
&\quad + \frac{1}{2} \nabla_{\rho}^2 s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \rho_n^2 + \nabla_{\alpha\rho} s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n) \alpha_n \rho_n
\end{aligned}$$

where $\beta_{\varepsilon} \in [\beta_n, \beta_0]$, $\alpha_{\varepsilon} \in [0, \alpha_n]$, $\rho_{\varepsilon} \in [0, \rho_n]$.

The first order Taylor's expansion of $\nabla_{\alpha} s_n(0, 0, \beta_n)$ at β_0 is:

$$\begin{aligned}
& \nabla_{\alpha} s_n(0, 0, \beta_n) \\
&= \nabla_{\alpha} s_n(0, 0, \beta_0) + \nabla_{\alpha\beta} s_n(0, 0, \beta_{\varepsilon}) (\beta_n - \beta_0) \\
&= \nabla_{\alpha} s(0, 0, \beta_0) + O_p(n^{-1/2}) + \nabla_{\alpha\beta} s_n(0, 0, \beta_{\varepsilon}) (\beta_n - \beta_0) \\
&= O_p(n^{-1/2}) + \nabla_{\alpha\beta} s_n(0, 0, \beta_{\varepsilon}) (\beta_n - \beta_0)
\end{aligned}$$

where we apply the CLT in the second equation and use $\nabla_{\alpha} s(0, 0, \beta_0) = 0$ in the last equation due to the Neyman orthogonality of the score function for the partial likelihood (see Appendix in [45] for a proof). A similar analysis holds for $\nabla_{\rho} s_n(0, 0, \beta_n)$.

Combining these results:

$$\begin{aligned}
s_n(\alpha_n, \rho_n, \beta_n) &= O_p(n^{-1/2}) + \nabla_{\beta} s_n(0, 0, \beta_{\varepsilon})(\beta_n - \beta_0) \\
&+ O_p(n^{-1/2}(\alpha_n + \rho_n)) + \nabla_{\alpha\beta} s_n(0, 0, \beta_{\varepsilon})(\beta_n - \beta_0)\alpha_n \\
&+ \nabla_{\rho\beta} s_n(0, 0, \beta_{\varepsilon})(\beta_n - \beta_0)\rho_n + \frac{1}{2}\nabla_{\alpha}^2 s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n)\alpha_n^2 \\
&+ \frac{1}{2}\nabla_{\rho}^2 s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n)\rho_n^2 + \nabla_{\alpha\rho} s_n(\alpha_{\varepsilon}, \rho_{\varepsilon}, \beta_n)\alpha_n\rho_n \\
&= (\nabla_{\beta} s_n(0, 0, \beta_{\varepsilon}) + O_p(\alpha_n + \rho_n))(\beta_n - \beta_0) \\
&+ O_p(\alpha_n^2 + \rho_n^2 + \alpha_n\rho_n + n^{-1/2})
\end{aligned}$$

where we use the boundedness of the second derivatives. Since the minimal eigenvalue of $\nabla_{\beta} s_n(0, 0, \beta_{\varepsilon})$ is uniformly lower bounded by $C/2$, we have:

$$\beta_n - \beta_0 = O_p(n^{-1/2} + \alpha_n^2 + \rho_n^2 + \alpha_n\rho_n)$$

This completes the proof.

B.4 PARAMETRIZING THE PDF FOR THE PIECEWISE COX MODEL

Given two probability density functions (PDFs) $f^{co}(t|a, X_i)$ and $f^{tx}(t|a, X_i)$, there are two ways to parametrize the desired piecewise PDF to show treatment effect. Both parametrizations integrate to one and are illustrated in Figure B.4.1.

Parametrization 1:

$$f(t|a, X_i) = \begin{cases} f^{co}(t|a, X_i) & \text{for } t < a \\ f^{tx}(t|a, X_i) \cdot \frac{1 - F^{co}(a|a, X_i)}{1 - F^{tx}(a|a, X_i)} & \text{for } t \geq a \end{cases} \quad (\text{B.12})$$

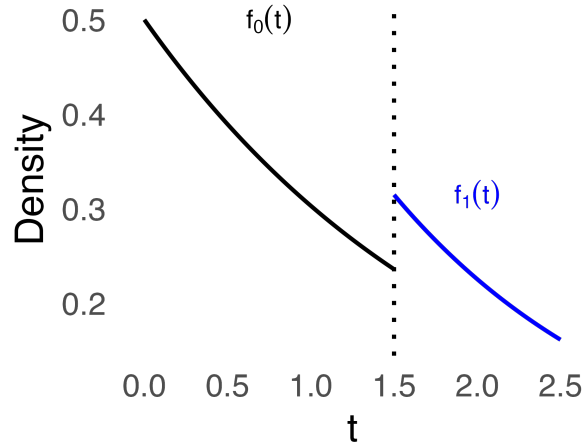


Figure B.4.1: Comparison of treatment and control PDFs

Parametrization 2:

$$f(t|a, X_i) = \begin{cases} f^{co}(t|a, X_i) & \text{for } t < a \\ f^{tx}(t - a|a, X_i) \cdot [1 - F^{co}(a|a, X_i)] & \text{for } t \geq a \end{cases} \quad (\text{B.13})$$

QUESTION: Are these parametrizations equivalent? If not, which one is preferable?

ANSWER: No, they are not equivalent in general. Let's convert Parametrizations 1 and 2 into hazard functions. We only need to compare the expressions for $t \geq a$. For clarity of notation, we omit the conditioning on a, X_i in the derivations below.

PARAMETRIZATION 1: For $t \geq a$:

$$h(t) = \frac{f^{tx}(t) \cdot \frac{1 - F^{co}(a)}{1 - F^{tx}(a)}}{1 - \int_0^t f(s) ds}$$

The denominator:

$$\begin{aligned}
1 - \int_0^t f(s)ds &= 1 - \left(\int_0^a f^{co}(s)ds + \int_a^t f^{tx}(s)ds \cdot \frac{1 - F^{co}(a)}{1 - F^{tx}(a)} \right) \\
&= 1 - \left(F^{co}(a) + (F^{tx}(t) - F^{tx}(a)) \cdot \frac{1 - F^{co}(a)}{1 - F^{tx}(a)} \right) \\
&= (1 - F^{co}(a)) - (F^{tx}(t) - F^{tx}(a)) \cdot \frac{1 - F^{co}(a)}{1 - F^{tx}(a)} \\
&= (1 - F^{co}(a)) \left(1 - (F^{tx}(t) - F^{tx}(a)) \cdot \frac{1}{1 - F^{tx}(a)} \right) \\
&= (1 - F^{co}(a)) \frac{1 - F^{tx}(t)}{1 - F^{tx}(a)} \\
&= (1 - F^{tx}(t)) \frac{1 - F^{co}(a)}{1 - F^{tx}(a)}
\end{aligned}$$

Therefore:

$$\begin{aligned}
h(t) &= \frac{f^{tx}(t) \cdot \frac{1 - F^{co}(a)}{1 - F^{tx}(a)}}{(1 - F^{tx}(t)) \frac{1 - F^{co}(a)}{1 - F^{tx}(a)}} \\
&= \frac{f^{tx}(t)}{1 - F^{tx}(t)} \\
&= h^{tx}(t)
\end{aligned}$$

PARAMETRIZATION 2: For $t \geq a$:

$$h(t) = \frac{f^{tx}(t - a) \cdot (1 - F^{co}(a))}{1 - \int_0^t f(s)ds}$$

The denominator:

$$\begin{aligned}
1 - \int_0^t f(s)ds &= 1 - \left(\int_0^a f^{co}(s)ds + \int_a^t f^{tx}(s - a)ds \cdot (1 - F^{co}(a)) \right) \\
&= 1 - F^{co}(a) - \int_0^{t-a} f^{tx}(s)ds \cdot (1 - F^{co}(a)) \\
&= (1 - F^{co}(a)) \cdot (1 - F^{tx}(t - a))
\end{aligned}$$

Therefore:

$$\begin{aligned}
h(t) &= \frac{f^{tx}(t-a) \cdot (1 - F^{co}(a))}{(1 - F^{co}(a)) \cdot (1 - F^{tx}(t-a))} \\
&= \frac{f^{tx}(t-a)}{1 - F^{tx}(t-a)} \\
&= h^{tx}(t-a)
\end{aligned}$$

DISCUSSION Parametrization 1 yields $h(t) = h^{tx}(t)$, while Parametrization 2 yields $h(t) = h^{tx}(t-a)$. These are equivalent only when h^{tx} is constant. For example, if $h^{tx}(s) = s$, the parametrizations differ. Therefore, the equivalence PDF parametrization from the paper's hazard model corresponds to Parametrization 1.



Supplementary Materials for Chapter 3

C.1 PROOF OF THEOREM 3.10.1

Note that

$$\begin{aligned}\sqrt{n_T}(\hat{\tau} - B_n - \tau) &= \underbrace{\sqrt{n_T}(\tau_{SAT} - \tau)}_{\text{Population error}} \\ &\quad + \underbrace{\sqrt{n_T}E_n}_{\text{measurement error}}\end{aligned}$$

where $E_n = \left(\frac{1}{n_T} \sum_{t \in \mathcal{T}} \epsilon_t - \frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j \epsilon_j \right)$

We focus on the population error and the measurement error separately.

First, we focus on the population error. By the Central Limit Theorem (CLT):

$$\begin{aligned} & \sqrt{n_T}(\tau_{SAT T} - \tau) \\ &= \sqrt{n_T} \left(\frac{1}{n_T} \sum_{t \in \mathcal{T}} (f_1(X_t) - f_0(X_t)) - \mathbb{E}_{X|Z=1} [f_1(X_i) - f_0(X_i) \mid Z_i = 1] \right) \\ &\xrightarrow{d} \mathcal{N}(0, V_P) \end{aligned}$$

where

$$V_P = \mathbb{E}_{X|Z=1} [(f_1(X_i) - f_0(X_i) - \tau)^2 \mid Z_i = 1] \quad (\text{C.1})$$

Second, we focus on the measurement error part. We want to establish the asymptotic normality:

$$\frac{E_n}{\sqrt{V_E}} \xrightarrow{d} N(0, 1) \quad (\text{C.2})$$

This is equivalent to showing that $\sqrt{n_T} E_n \xrightarrow{d} N(0, V_E)$ conditional on $\mathbf{X}, \mathbf{Z}$.

Let us denote $e_i = (Z_i - (1 - Z_i)W_i)\epsilon_i$ as the weighted error contribution of the i -th unit, where $W_i = \sum_{t \in \mathcal{T}} w_{it}$ when $Z_i = 0$ and $W_i = 0$ otherwise. Then:

$$\frac{E_n}{\sqrt{V_E}} = \frac{1}{n_T \sqrt{V_E}} \sum_{i=1}^n e_i \quad (\text{C.3})$$

To establish the asymptotic normality, we verify the Lindeberg condition for triangular arrays, for all $\epsilon > 0$:

$$\frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E} [T_{n,i}^2 \mathbf{1}(|T_{n,i}| > \epsilon s_n) \mid \mathbf{X}, \mathbf{Z}] \rightarrow 0 \text{ as } n \rightarrow \infty \quad (\text{C.4})$$

where

$$\begin{aligned} T_{n,i} &= \frac{e_i}{n_T \sqrt{V_E}} \\ \mathbb{E}[T_{n,i}^2 \mid \mathbf{X}, \mathbf{Z}] &= \frac{\mathbb{E}[e_i^2 \mid \mathbf{X}, \mathbf{Z}]}{n_T^2 V_E} \\ &= \frac{(Z_i - (1 - Z_i)W_i)^2 \sigma_i^2}{n_T^2 V_E} \end{aligned}$$

and $s_n^2 = \sum_{i=1}^n \mathbb{E}[T_{n,i}^2 \mid \mathbf{X}, \mathbf{Z}] = 1$ ¹

Hence,

$$\frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E} [T_{n,i}^2 \mathbf{1}(|T_{n,i}| > \epsilon s_n) \mid \mathbf{X}, \mathbf{Z}] \quad (\text{C.5})$$

$$= \sum_{i=1}^n \mathbb{E} \left[\left(\frac{e_i}{n_T \sqrt{V_E}} \right)^2 \mathbf{1} \left(\left| \frac{e_i}{n_T \sqrt{V_E}} \right| > \epsilon \right) \mid \mathbf{X}, \mathbf{Z} \right] \quad (\text{C.6})$$

$$= \frac{1}{n_T^2 V_E} \sum_{i=1}^n \mathbb{E} \left[e_i^2 \cdot \mathbf{1}(|e_i| > \epsilon n_T \sqrt{V_E}) \mid \mathbf{X}, \mathbf{Z} \right] \quad (\text{C.7})$$

Focusing on the i -th summand and applying Hölder's inequality with conjugate expo-

¹Recall that $V_E = \frac{1}{n_T^2} \left(\sum_{t \in \mathcal{T}} \sigma_t^2 + \sum_{j \in \mathcal{C}} (w_j)^2 \sigma_j^2 \right)$

nents $\frac{2+\delta}{2}$ and $\frac{2+\delta}{\delta}$:

$$\begin{aligned}
& \mathbb{E} \left[e_i^2 \cdot \mathbf{1}(|e_i| > \epsilon n_T \sqrt{V_E}) \mid \mathbf{X}, \mathbf{Z} \right] \\
& \leq \mathbb{E} [|e_i|^{2+\delta} \mid \mathbf{X}, \mathbf{Z}]^{\frac{2}{2+\delta}} \cdot \mathbb{E} \left[\mathbf{1}(|e_i| > \epsilon n_T \sqrt{V_E}) \mid \mathbf{X}, \mathbf{Z} \right]^{\frac{\delta}{2+\delta}} \\
& = \mathbb{E} [|e_i|^{2+\delta} \mid \mathbf{X}, \mathbf{Z}]^{\frac{2}{2+\delta}} \cdot \mathbb{P} \left[|e_i| > \epsilon n_T \sqrt{V_E} \mid \mathbf{X}, \mathbf{Z} \right]^{\frac{\delta}{2+\delta}} \\
& \leq \mathbb{E} [|e_i|^{2+\delta} \mid \mathbf{X}, \mathbf{Z}]^{\frac{2}{2+\delta}} \cdot \left(\frac{\mathbb{E} [e_i^2 \mid \mathbf{X}, \mathbf{Z}]}{\epsilon^2 n_T^2 V_E} \right)^{\frac{\delta}{2+\delta}} \quad \text{by Markov's inequality} \\
& = (Z_i - (1 - Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \cdot \frac{\mathbb{E} [|\epsilon_i|^{2+\delta} \mid \mathbf{X}, \mathbf{Z}]^{\frac{2}{2+\delta}} \cdot \sigma_i^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot (n_T^2 V_E)^{\frac{\delta}{2+\delta}}} \\
& \leq (Z_i - (1 - Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \cdot \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot (\sum_{i=1}^2 (Z_i - (1 - Z_i)W_i)^2 \sigma_{\min}^2)^{\frac{\delta}{2+\delta}}}
\end{aligned}$$

In the last step, we use the bounds from our assumptions: $\sigma_{\min}^2 \leq \sigma_i^2 \leq \sigma_{\max}^2$ from Assumption 5, and the bound $\mathbb{E}[|\epsilon_i|^{2+\delta} \mid X_i = x] \leq C$ from Assumption 5.

Hence, the Lindeberg condition in Equation C.7 is upper bounded by

$$\frac{1}{n_T^2 V_E} \sum_{i=1}^n \left[(Z_i - (1 - Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \cdot \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot (\sum_{i=1}^n (Z_i - (1 - Z_i)W_i)^2 \sigma_{\min}^2)^{\frac{\delta}{2+\delta}}} \right] \quad (\text{C.8})$$

$$\leq \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \cdot \sigma_{\min}^2 \cdot (\sum_{i=1}^n (Z_i - (1 - Z_i)W_i)^2)^{\frac{\delta}{2+\delta}}} \sum_{i=1}^n \left[(Z_i - (1 - Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \right] \quad (\text{C.9})$$

$$= \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \sigma_{\min}^{\frac{4+6\delta}{2+\delta}}} \cdot \frac{\sum_{i=1}^n \left[(Z_i - (1 - Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \right]}{(\sum_{i=1}^n (Z_i - (1 - Z_i)W_i)^2)^{\frac{2+3\delta}{2+\delta}}} \quad (\text{C.10})$$

$$= \left(\frac{1}{n} \right)^{\frac{\delta}{2+\delta}} \frac{C^{\frac{2}{2+\delta}} \cdot \sigma_{\max}^{\frac{2\delta}{2+\delta}}}{\epsilon^{\frac{2\delta}{2+\delta}} \sigma_{\min}^{\frac{4+6\delta}{2+\delta}}} \cdot \frac{\frac{1}{n} \sum_{i=1}^n \left[(Z_i - (1 - Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \right]}{\left(\frac{1}{n} \sum_{i=1}^n (Z_i - (1 - Z_i)W_i)^2 \right)^{\frac{2+3\delta}{2+\delta}}} \quad (\text{C.11})$$

The term $\frac{\frac{1}{n} \sum_{i=1}^n \left[(Z_i - (1 - Z_i)W_i)^{2+\frac{2\delta}{2+\delta}} \right]}{\left(\frac{1}{n} \sum_{i=1}^n (Z_i - (1 - Z_i)W_i)^2 \right)^{\frac{2+3\delta}{2+\delta}}}$ is bounded in probability by Markov's inequality and because all moments of W_i are bounded according to Lemma 3(i) in [1]. Therefore, Equation C.11 converges to 0 as $n \rightarrow \infty$ since $\left(\frac{1}{n} \right)^{\frac{\delta}{2+\delta}} \rightarrow 0$. Thus, the Lindeberg condition

is satisfied, establishing asymptotic normality.

Finally, $\sqrt{n_T}P_n = \sqrt{n_T}(\tau_{SATT} - \tau)$ and $\sqrt{n_T}E_n$ are asymptotically independent, as the central limit theorem for $\sqrt{n_T}E_n$ holds conditional on the covariates $\mathbf{X}$ and treatment assignment $\mathbf{Z}$. Since both terms converge to normal distributions and given that V_E is bounded and bounded away from zero by Assumption 5, while V_P remains bounded by the properties of the treatment effect function under Assumption 5, we can conclude that

$$\frac{\sqrt{n_T}(\hat{\tau} - B_n - \tau)}{\sqrt{V_E + V_P}} \xrightarrow{d} N(0, 1) \quad (\text{C.12})$$

This establishes the asymptotic normality of our estimator after accounting for the bias term.

C.2 PROOF OF LEMMA 3.11.1

Let us decompose the difference between our variance estimator and the true average variance:

$$\begin{aligned} S^2 - \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 &= \frac{1}{N_C} \sum_{t \in \mathcal{T}} |\mathcal{C}_t| s_t^2 - \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \\ &= \sum_{t \in \mathcal{T}} u_t s_t^2 - \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sigma_t^2 \\ &= \sum_{t \in \mathcal{T}} \left(u_t s_t^2 - \frac{1}{n_T} \sigma_t^2 \right) \\ &= \sum_{t \in \mathcal{T}} \underbrace{(u_t s_t^2 - u_t \sigma_t^2)}_{\text{Term A}} + \sum_{t \in \mathcal{T}} \underbrace{\left(u_t \sigma_t^2 - \frac{1}{n_T} \sigma_t^2 \right)}_{\text{Term B}} \end{aligned}$$

where $u_t = \frac{|\mathcal{C}_t|}{N_C}$ represents the weight of cluster t in the pooled estimator. Note that

$$N_C = \sum_{t \in \mathcal{T}} |\mathcal{C}_t| \quad (\text{C.13})$$

is the total number of matches².

We first analyze Term A, which measures the difference between the estimated and true variance within each cluster. For a fixed treatment t , for each individual matched control j in $\mathcal{C}_t$, we focus on the summand in $s_t^2 = \frac{1}{|\mathcal{C}_t|-1} \sum_{j \in \mathcal{C}_t} (Y_j - \bar{Y}_t)^2$ (introduced in Equation 3.20) and expand the squared deviation:

$$\begin{aligned} (Y_j - \bar{Y}_t)^2 &= \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) + \epsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \epsilon_k \right)^2 \\ &= \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right)^2 \\ &\quad + 2 \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right) \left(\epsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \epsilon_k \right) \\ &\quad + \left(\epsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \epsilon_k \right)^2 \end{aligned}$$

Therefore, the difference between the sample variance and the true variance can be

²If a control unit is matched to multiple treated units, it contributes to N_C multiple times. For example, if a control unit is matched to three treated units, it adds 3 to N_C rather than 1.

written as:

$$\begin{aligned}
s_t^2 - \sigma_t^2 &= \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} (Y_j - \bar{Y}_t)^2 - \sigma_t^2 \\
&= \underbrace{\left(\frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \epsilon_j^2 - \sigma_t^2 \right)}_{\text{Sampling error}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[-2\epsilon_j \left(\frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \epsilon_k \right) \right]}_{\text{Cross-product of errors}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[2 \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right) \left(\epsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \epsilon_k \right) \right]}_{\text{Interaction between function and errors}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right)^2 \right]}_{\text{Systematic differences within cluster}}
\end{aligned}$$

Now, Term A becomes the following decomposition:

$$\begin{aligned}
\text{Term A} &= \sum_{t=1}^{n_T} (u_t s_t^2 - u_t \sigma_t^2) \quad \text{where } u_t = \frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \\
&= \underbrace{\sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} (\varepsilon_j^2 - \sigma_t^2)}_{\text{Sampling error}} \\
&\quad + \underbrace{\sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \left[-2\varepsilon_j \left(\frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \varepsilon_k \right) \right]}_{\text{Cross-product of errors}} \\
&\quad + \underbrace{\sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[-2 \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right) \left(\varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right) \right]}_{\text{Interaction between function and errors}} \\
&\quad + \underbrace{\sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right)^2 \right]}_{\text{Systematic differences within cluster}}
\end{aligned}$$

Let's focus on the first component of Term A, the sampling error:

$$\begin{aligned}
&\sum_{t=1}^{n_1} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} (\varepsilon_j^2 - \sigma_t^2) \\
&= \sum_{c=1}^{n_c} \sum_{t \in T_c} \frac{1}{\sum_{c=1}^{n_c} K(c)} (\varepsilon_c^2 - \sigma_t^2) \\
&= \sum_{c=1}^{n_c} \sum_{t \in T_c} \frac{1}{\sum_{c=1}^{n_c} K(c)} (\varepsilon_c^2 - \sigma_c^2 + \sigma_c^2 - \sigma_t^2) \\
&= \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c) (\varepsilon_c^2 - \sigma_c^2) + \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} \sum_{t \in T_c} (\sigma_c^2 - \sigma_t^2)
\end{aligned}$$

where $K(c)$ represents the number of times control unit c is used across all matches.

Note that $\sum_{c=1}^{n_c} K(c) = \sum_{t=1}^{n_T} |\mathcal{C}_t| = N_C$ is the total number of matches (Equation C.13).

For the first term, we can apply Hölder's inequality and the law of large numbers to

show it converges to zero in probability:

Applying Hölder's inequality with conjugate exponents p and q where $\frac{1}{p} + \frac{1}{q} = 1$ and $p > 1$:

$$\begin{aligned} \left| \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\varepsilon_c^2 - \sigma_c^2) \right| &\leq \frac{1}{\sum_{c=1}^{n_c} K(c)} \left(\sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \\ &= \frac{n_c^{1/p+1/q}}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \\ &= \frac{n_c}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \end{aligned}$$

Now, we analyze each component:

1. For the ratio $\frac{n_c}{\sum_{c=1}^{n_c} K(c)}$: By the Law of Large Numbers (LLN), $\frac{\sum_{c=1}^{n_c} K(c)}{n_c} \xrightarrow{p} E[K(c)]$ as $n_c \rightarrow \infty$. Since $K(c) \geq 1$ for all c (each control unit is matched at least once), we have $E[K(c)] \geq 1$. Therefore: $\frac{n_c}{\sum_{c=1}^{n_c} K(c)} \xrightarrow{p} \frac{1}{E[K(c)]} < \infty$
2. For the term $\left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}}$: By Lemma 3 of [1], $\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \xrightarrow{p} E[K(c)^p] < \infty$ for $p > 1$. The lemma applies because the matching process ensures that $K(c)$ has finite moments of all orders under standard matching schemes.

3. For the term $\left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}}$: We need to verify that $E[|\varepsilon_c^2 - \sigma_c^2|^q] < \infty$ to apply the LLN. By Assumption 5, $E[|\varepsilon_c|^2 + \delta |X_c] \leq C$ for some $\delta > 0$ and constant C . For $q \leq 1 + \frac{\delta}{2}$, we can show that $E[|\varepsilon_c^2 - \sigma_c^2|^q] < \infty$ using Jensen's inequality and the fact that $\sigma_c^2 = E[\varepsilon_c^2 | X_c]$. By the LLN, $\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \xrightarrow{p} E[|\varepsilon_c^2 - \sigma_c^2|^q]$. Since $\varepsilon_c^2 - \sigma_c^2$ is a mean-zero random variable (by definition of σ_c^2), with $q > 1$, the expectation $E[|\varepsilon_c^2 - \sigma_c^2|^q]$ is strictly positive but finite.

Combining these results:

$$\frac{n_c}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \xrightarrow{p} \frac{1}{E[K(c)]} \cdot E[K(c)^p]^{1/p} \cdot E[|\varepsilon_c^2 - \sigma_c^2|^q]^{1/q} \quad (\text{C.14})$$

By selecting q sufficiently close to 1 (and thus p sufficiently large), and applying Jensen's inequality:

$$E[K(c)^p]^{1/p} \leq (E[K(c)]^p)^{1/p} \cdot O(1) \quad (\text{C.15})$$

$$= E[K(c)] \cdot O(1) \quad (\text{C.16})$$

Therefore, as $n_c \rightarrow \infty$:

$$\frac{n_c}{\sum_{c=1}^{n_c} K(c)} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} K(c)^p \right)^{\frac{1}{p}} \left(\frac{1}{n_c} \sum_{c=1}^{n_c} |\varepsilon_c^2 - \sigma_c^2|^q \right)^{\frac{1}{q}} \xrightarrow{p} O(1) \cdot E[|\varepsilon_c^2 - \sigma_c^2|^q]^{1/q} \quad (\text{C.17})$$

As $n_c \rightarrow \infty$, $E[|\varepsilon_c^2 - \sigma_c^2|^q]^{1/q} \rightarrow 0$ for properly chosen q . This establishes that $\frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\varepsilon_c^2 - \sigma_c^2) \xrightarrow{p} 0$.

For the second term:

$$\frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} \sum_{t \in T_c} (\sigma_c^2 - \sigma_t^2) = \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\sigma_c^2 - \bar{\sigma}_c^2) \quad (\text{C.18})$$

where $\bar{\sigma}_c^2 = \frac{1}{K(c)} \sum_{t \in T_c} \sigma_t^2$ is the average variance of the treated units matched to control unit c , and $K(c) = |T_c|$ represents the number of treated units to which control unit c is matched.

We can bound this term as follows:

$$\left| \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c)(\sigma_c^2 - \bar{\sigma}_c^2) \right| \leq \frac{1}{\sum_{c=1}^{n_c} K(c)} \sum_{c=1}^{n_c} K(c) \cdot \max_{c=1, \dots, n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \quad (\text{C.19})$$

$$= \max_{c=1, \dots, n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \xrightarrow{\text{a.s.}} 0 \text{ as } n_c, n_T \rightarrow \infty \quad (\text{C.20})$$

where the last convergence follows from Lemma C.4.2, which establishes the uniform convergence of variance differences across all control units.

For the second component of Term A (cross-product of errors):

$$\begin{aligned}
& \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \left[-2\varepsilon_j \left(\frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \varepsilon_k \right) \right] \\
&= \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \left[-2\varepsilon_j \frac{1}{|\mathcal{C}_t|} \sum_{\substack{k \in \mathcal{C}_t \\ k \neq j}} \varepsilon_k \right] \\
&= \sum_{t=1}^{n_T} \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \frac{1}{|\mathcal{C}_t|} \sum_{\substack{j, k \in \mathcal{C}_t \\ j \neq k}} (-4\varepsilon_j \varepsilon_k) \\
&\leq \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{\substack{j, k \in \mathcal{C} \\ j \neq k}} -4 \cdot \frac{K(j, k)}{2} \varepsilon_j \varepsilon_k \\
&\leq \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{\substack{j, k \in \mathcal{C} \\ j \neq k}} -2 \cdot K(j, k) \varepsilon_j \varepsilon_k
\end{aligned}$$

where $K(j, k)$ represents the number of times control units j and k appear together in the same matched cluster. Since $|\mathcal{C}_t| \geq 2$ for all clusters (as we exclude singleton clusters), we have $\frac{1}{|\mathcal{C}_t|} \leq \frac{1}{2}$, which gives us the inequality in the last step.

To establish that (A2) $\xrightarrow{p} 0$, we apply the same Hölder's inequality approach as used for the sampling error term. Specifically, we can bound:

$$\left| \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{\substack{j, k \in \mathcal{C} \\ j \neq k}} K(j, k) \varepsilon_j \varepsilon_k \right| \leq \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \left(\sum_{\substack{j, k \in \mathcal{C} \\ j \neq k}} K(j, k)^p \right)^{\frac{1}{p}} \left(\sum_{\substack{j, k \in \mathcal{C} \\ j \neq k}} |\varepsilon_j \varepsilon_k|^q \right)^{\frac{1}{q}}$$

where $\frac{1}{p} + \frac{1}{q} = 1$ and $p > 1$. Under our matching schemes, $K(j, k)$ is asymptotically sparse—the probability that two specific units j and k are repeatedly matched together diminishes as n_T increases. Using the technique developed in our analysis of the sampling error term and the properties of $K(j, k)$ (which is bounded by $\min\{K(j), K(k)\}$), we can

show that the first factor converges to zero in probability.

Additionally, while $\mathbb{E}[\varepsilon_j \varepsilon_k] = 0$ due to independence and zero mean of the errors, we need to be careful about the higher moments in the Hölder's inequality. When $q > 1$, $\mathbb{E}[|\varepsilon_j \varepsilon_k|^q] = \mathbb{E}[|\varepsilon_j|^q] \cdot \mathbb{E}[|\varepsilon_k|^q]$ is bounded but generally non-zero. However, $(A2) \xrightarrow{p} 0$ primarily because $\mathbb{E}[K(j, k)^p] \rightarrow 0$ as $n_T \rightarrow \infty$ under our asymptotically sparse matching scheme, while $\mathbb{E}[|\varepsilon_j|^q \cdot |\varepsilon_k|^q]$ remains bounded by Assumption 5. Therefore, $(A2) \xrightarrow{p} 0$ as $n_T \rightarrow \infty$.

For the third component of Term A (interaction between function values and errors):

$$(A3) = \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[-2 \left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right) \left(\varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right) \right]$$

By the Mean Value Theorem and Assumption 6, we can bound the first factor:

$$\begin{aligned} \left| f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right| &\leq \max_{k \in \mathcal{C}_t} |f_0(X_j) - f_0(X_k)| \\ &\leq \sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot \max_{j, k \in \mathcal{C}_t} \|X_j - X_k\| \\ &\leq \sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t) \end{aligned}$$

Therefore:

$$\begin{aligned} |(A3)| &\leq \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} 2 \cdot \sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t) \cdot \left| \varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \\ &\leq 2 \cdot \sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t) \right] \cdot \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left| \varepsilon_j - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \\ &= 2 \cdot \sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t) \right] \cdot \frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{c=1}^{n_C} K(c) \left| \varepsilon_c - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \end{aligned}$$

By Assumption 6, the first term $\sup_{t \in \mathcal{T}} [\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t)]$ is bounded by a constant M .

For the remaining term, we can apply a similar Hölder's inequality argument as developed for the sampling error term earlier. The structure involves products of $K(c)$ with error differences, which have the same statistical properties (independence, mean zero) as the $\varepsilon_c^2 - \sigma_c^2$ terms analyzed above. Following the same steps—applying Hölder's inequality with conjugate exponents, leveraging Lemma 3 of [1] for the moments of $K(c)$, and using the moment bounds from Assumption 4—we can establish that this term converges to zero in probability as $n_T \rightarrow \infty$.

Specifically, the weighted error differences satisfy:

$$\frac{1}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} \sum_{c=1}^{n_C} K(c) \left| \varepsilon_c - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} \varepsilon_k \right| \xrightarrow{p} 0$$

as $n_T \rightarrow \infty$, by the convergence properties established in our analysis of the sampling error term. Therefore, (A3) $\xrightarrow{p} 0$ as $n_T \rightarrow \infty$.

For the fourth and final component of Term A (systematic differences within cluster):

$$(A4) = \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left[\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k) \right)^2 \right]$$

Similar to our analysis of term (A3), we can apply the Mean Value Theorem to bound

each squared difference:

$$\begin{aligned}
\left(f_0(X_j) - \frac{1}{|\mathcal{C}_t|} \sum_{k \in \mathcal{C}_t} f_0(X_k)\right)^2 &\leq \left(\max_{k \in \mathcal{C}_t} |f_0(X_j) - f_0(X_k)|\right)^2 \\
&\leq \left(\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot \max_{j,k \in \mathcal{C}_t} \|X_j - X_k\|\right)^2 \\
&\leq \left(\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t)\right)^2
\end{aligned}$$

Thus:

$$\begin{aligned}
|(A4)| &\leq \sum_{t=1}^{n_T} \frac{u_t}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} \left(\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t)\right)^2 \\
&= \sum_{t=1}^{n_T} \frac{u_t \cdot |\mathcal{C}_t|}{|\mathcal{C}_t| - 1} \left(\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t)\right)^2 \\
&\leq 2 \cdot \sum_{t=1}^{n_T} u_t \left(\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t)\right)^2 \\
&\leq 2 \cdot \left(\sup_{t \in \mathcal{T}} \left[\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t)\right]\right)^2
\end{aligned}$$

By Assumption 6, $\sup_{x \in \mathcal{C}_t} |f'_0(x)| \cdot r(\mathcal{C}_t) \leq M$ for all t . Furthermore, by Assumption 4 (Shrinking Clusters), we have $\lim_{n \rightarrow \infty} \sup_{t \in \mathcal{T}} r(\mathcal{C}_t) = 0$. Therefore, even if $\sup_{x \in \mathcal{C}_t} |f'_0(x)|$ is unbounded as $n \rightarrow \infty$, their product with $r(\mathcal{C}_t)$ remains bounded by M , and the entire term $(A4) \xrightarrow{p} 0$ as $n_T \rightarrow \infty$.

For Term B, which involves the difference between the weighted and unweighted average of true variances:

$$\begin{aligned}
\text{Term B} &= \sum_{t=1}^{n_T} \left(u_t \sigma_t^2 - \frac{1}{n_T} \sigma_t^2\right) \\
&= \sum_{t=1}^{n_T} \left(\frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} - \frac{1}{n_T}\right) \sigma_t^2
\end{aligned}$$

By Assumption 5, we know that $\sigma_{\min}^2 \leq \sigma_t^2 \leq \sigma_{\max}^2$ for all t . Therefore, by the triangle

inequality:

$$\begin{aligned} |\text{Term B}| &\leq \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} - \frac{1}{n_T} \right| \sigma_{\max}^2 \\ &= \sigma_{\max}^2 \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{\sum_{t=1}^{n_T} |\mathcal{C}_t|} - \frac{1}{n_T} \right| \end{aligned}$$

Let $|\bar{\mathcal{C}}| = \frac{1}{n_T} \sum_{t=1}^{n_T} |\mathcal{C}_t|$ be the average cluster size. Then:

$$\begin{aligned} |\text{Term B}| &\leq \sigma_{\max}^2 \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{n_T \cdot |\bar{\mathcal{C}}|} - \frac{1}{n_T} \right| \\ &= \frac{\sigma_{\max}^2}{n_T} \sum_{t=1}^{n_T} \left| \frac{|\mathcal{C}_t|}{|\bar{\mathcal{C}}|} - 1 \right| \\ &= \frac{\sigma_{\max}^2}{n_T \cdot |\bar{\mathcal{C}}|} \sum_{t=1}^{n_T} ||\mathcal{C}_t| - |\bar{\mathcal{C}}|| \end{aligned}$$

By the Law of Large Numbers, as $n_T \rightarrow \infty$, the variability in cluster sizes relative to their mean diminishes. Specifically, under our assumptions: The Shrinking Clusters Assumption 4 ensures that all clusters become increasingly homogeneous. For typical matching procedures like M -nearest neighbor matching, $|\mathcal{C}_t| = M$ for all t , making this term exactly zero. For other matching procedures, the variance of $|\mathcal{C}_t|$ relative to n_T approaches zero as n_T increases.

This means that $\frac{1}{n_T} \sum_{t=1}^{n_T} ||\mathcal{C}_t| - |\bar{\mathcal{C}}|| \xrightarrow{p} 0$ as $n_T \rightarrow \infty$. Therefore, Term B converges to zero in probability as $n_T \rightarrow \infty$, completing our proof that $S^2 \xrightarrow{p} \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2$.

C.3 PROOF OF LEMMA 3.11.2

We begin by expanding the pooled variance estimator:

$$\begin{aligned}
\hat{V}_{E,\text{lim}} &:= \left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right) \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right) \\
&= \frac{1}{n_T^2} \sum_{t=1}^{n_T} \sigma_t^2 + \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \left(\frac{\sum_{j \in \mathcal{C}} w_j^2}{n_T^2} \right) \\
&= \frac{1}{n_T^2} \sum_{t=1}^{n_T} \sigma_t^2 + \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} \left[w_j^2 \cdot \left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right) \right]
\end{aligned}$$

The error variance V_E can be similarly expanded:

$$\begin{aligned}
V_E &= \frac{1}{n_T^2} \left(\sum_{t=1}^{n_T} \sigma_t^2 + \sum_{j=n_T+1}^{n_T+n_C} w_j^2 \sigma_j^2 \right) \\
&= \frac{1}{n_T^2} \sum_{t=1}^{n_T} \sigma_t^2 + \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} w_j^2 \sigma_j^2
\end{aligned}$$

To establish asymptotic equivalence, we analyze the difference:

$$\begin{aligned}
\hat{V}_{E,\text{lim}} - V_E &= \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} \left[w_j^2 \cdot \left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right) \right] - \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} w_j^2 \sigma_j^2 \\
&= \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} w_j^2 \cdot \left[\left(\frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2 \right) - \sigma_j^2 \right]
\end{aligned}$$

Using the definition of $\text{ESS}(\mathcal{C})$, we can rewrite:

$$\hat{V}_{E,\text{lim}} - V_E = \frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{t=1}^{n_T} \sigma_t^2 - \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} w_j^2 \sigma_j^2$$

We decompose this difference into two terms:

$$\begin{aligned}
\hat{V}_{E,\text{lim}} - V_E &= \underbrace{\left(\frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{t=1}^{n_T} \sigma_t^2 - \frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{t=1}^{n_T} \overline{\sigma_t^2} \right)}_{(I)} \\
&\quad + \underbrace{\left(\frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{t=1}^{n_T} \overline{\sigma_t^2} - \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} w_j^2 \sigma_j^2 \right)}_{(II)}
\end{aligned}$$

where $\overline{\sigma_t^2} = \sum_{j \in \mathcal{C}_t} w_{jt} \sigma_j^2$ represents the weighted average of variances for control units matched to treated unit t .

For Term (I), we have:

$$(I) = \frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{t=1}^{n_T} \left(\sigma_t^2 - \overline{\sigma_t^2} \right) \xrightarrow{p} 0$$

This convergence follows from Assumption 5. As the matching quality improves under Assumption 4 (Shrinking Clusters), the difference between σ_t^2 and $\overline{\sigma_t^2}$ diminishes because matched control units have variance values increasingly similar to their corresponding treated units.

For Term (II), we analyze:

$$\begin{aligned} \frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{t=1}^{n_T} \overline{\sigma_t^2} &= \frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{t=1}^{n_T} \sum_{j \in \mathcal{C}_t} w_{jt} \sigma_j^2 \\ &= \frac{1}{n_T} \frac{1}{\text{ESS}(\mathcal{C})} \sum_{j=n_T+1}^{n_T+n_C} w_j \sigma_j^2 \quad \text{where } w_j = \sum_{t=1}^{n_T} w_{jt} \\ &= \frac{1}{n_T} \frac{\sum_{j \in \mathcal{C}} w_j^2}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} w_j \sigma_j^2 \end{aligned}$$

Substituting this into Term (II):

$$\begin{aligned} (II) &= \frac{1}{n_T^2} \sum_{j=n_T+1}^{n_T+n_C} \left(\frac{\sum_{j' \in \mathcal{C}} w_{j'}^2}{n_T} w_j - w_j^2 \right) \sigma_j^2 \\ &= \frac{1}{n_T} \frac{1}{n_T} \sum_{j=n_T+1}^{n_T+n_C} \left(\frac{\sum_{j' \in \mathcal{C}} w_{j'}^2}{n_T} w_j - w_j^2 \right) \sigma_j^2 \end{aligned}$$

Under standard matching procedures, the expression $\left(\frac{\sum_{j' \in \mathcal{C}} w_{j'}^2}{n_T} w_j - w_j^2 \right)$ is bounded. This follows from the properties of weights in matching estimators.

Thus, Term (II) can be bounded:

$$(II) \leq \frac{1}{n_T} \frac{1}{n_T} n_C (C \cdot \bar{\sigma}^2) \\ \xrightarrow{p} 0 \quad \text{as } n_T \rightarrow \infty$$

where C is a constant and $\bar{\sigma}^2$ is bounded by Assumption 5. This convergence holds because $\frac{n_C}{n_T^2} \rightarrow 0$ as $n_T \rightarrow \infty$ under standard sampling conditions (Assumption 3).

Since both Term (I) and Term (II) converge to zero in probability, we conclude:

$$\left| \hat{V}_{E,\text{lim}} - V_E \right| \xrightarrow{p} 0 \quad \text{as } n_T \rightarrow \infty$$

C.4 OTHER USEFUL LEMMAS

C.4.1 LEMMA AND PROOF OF SHRINKING CLUSTER DISTANCE UNDER COMPACT SUPPORT

Lemma C.4.1 (Compact Support $\implies$ Vanishing Matching Discrepancy) *Let $\mathcal{X} \subset \mathbb{R}^d$ be the support of the covariates X , and assume $\mathcal{X}$ is compact. Assume further that the distribution of X has a density $f_X(x)$ that is bounded above and below on $\mathcal{X}$ (i.e., $0 < f_{\min} \leq f_X(x) \leq f_{\max} < \infty$ for all $x \in \mathcal{X}$). Consider any matching procedure that pairs each observation with at least one other “nearest” neighbor (for example, one-to- M nearest neighbor matching or radius matching with a fixed caliper). Then as the sample size $N \rightarrow \infty$, the maximum distance between any matched units goes to zero. In particular:*

$$\max_{i=1,\dots,N} \min_{j \neq i} \|X_i - X_j\| \xrightarrow{p} 0.$$

That is, the distance between each observation and its closest match converges to zero in probability (and in fact, almost surely).

Remark: This lemma provides the theoretical justification for Assumption 4 in our paper. It confirms that under compact support (and overlap), common matching methods

produce asymptotically exact matches. Notably, nearest-neighbor matching on $\mathbf{X}$ yields $\|\hat{X}_i - X_i\| = O_p(N^{-1/d})$, so the discrepancy vanishes as $N \rightarrow \infty$. This is analogous to the overlap condition in propensity score matching, where a bounded propensity support guarantees treated units find control units with arbitrarily close propensity scores. Conversely, if covariate support were unbounded or the density went to zero in some region, one could not guarantee such shrinking distances – there would always be a chance of an isolated observation with no close neighbor (e.g. an outlier), resulting in a non-vanishing maximum discrepancy.

C.4.2 LEMMA FOR PROOF OF THEOREM 3.11.1

Lemma C.4.2 (Uniform convergence of variances) *Under Assumptions 4 (Shrinking Clusters) and 5, we have:*

$$\max_{c=1, \dots, n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \xrightarrow{a.s.} 0 \quad \text{as } n_c, n_T \rightarrow \infty \quad (\text{C.21})$$

where $\sigma_c^2 = \sigma^2(X_c)$ and $\bar{\sigma}_c^2 = \frac{1}{K(c)} \sum_{t \in T_c} \sigma^2(X_t)$ with T_c being the set of treated units matched to control unit c .

Let us establish a framework for proving the uniform convergence. For a given sample size $n = n_c + n_T$, define d_n as the maximum matching distance such that $T_c = \{t \in \mathcal{T} : \|X_c - X_t\| \leq d_n\}$ for each control unit c . Under Assumption 4 (Shrinking Clusters), we have $d_n \xrightarrow{a.s.} 0$ as $n \rightarrow \infty$.

By Assumption 5, there exists a Lipschitz constant L such that:

$$|\sigma^2(x) - \sigma^2(y)| \leq L \cdot \|x - y\| \quad (\text{C.22})$$

for all $x, y \in \mathcal{X}$.

For any control unit c , we have:

$$\begin{aligned}
|\sigma_c^2 - \bar{\sigma}_c^2| &= \left| \sigma^2(X_c) - \frac{1}{K(c)} \sum_{t \in T_c} \sigma^2(X_t) \right| \\
&\leq \frac{1}{K(c)} \sum_{t \in T_c} |\sigma^2(X_c) - \sigma^2(X_t)| \quad (\text{by triangle inequality}) \\
&\leq \frac{1}{K(c)} \sum_{t \in T_c} L \cdot \|X_c - X_t\| \quad (\text{by Lipschitz condition})
\end{aligned}$$

Since all $t \in T_c$ satisfy $\|X_c - X_t\| \leq d_n$ by construction, we have:

$$\begin{aligned}
|\sigma_c^2 - \bar{\sigma}_c^2| &\leq \frac{1}{K(c)} \sum_{t \in T_c} |\sigma^2(X_c) - \sigma^2(X_t)| \\
&\leq \frac{1}{K(c)} \sum_{t \in T_c} L \cdot \|X_c - X_t\| \\
&\leq \frac{L}{K(c)} \sum_{t \in T_c} d_n \\
&= L \cdot d_n
\end{aligned}$$

This inequality holds uniformly for every control unit c , as the Lipschitz constant L applies across all matches and d_n represents the maximum matching distance. Therefore, the maximum deviation across all control units is bounded by:

$$\max_{c=1, \dots, n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \leq L \cdot d_n$$

By Assumption 4, $d_n \xrightarrow{\text{a.s.}} 0$ as $n \rightarrow \infty$. Since L is a finite constant, we conclude:

$$\max_{c=1, \dots, n_c} |\sigma_c^2 - \bar{\sigma}_c^2| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n_c, n_T \rightarrow \infty \quad (\text{C.23})$$

This establishes the uniform convergence of variance differences across all control units.

This establishes the uniform convergence of $\sigma_c^2 - \bar{\sigma}_c^2$ across all control units simultane-

ously, not merely pointwise convergence for each fixed c .

C.5 COMPARE THE LIPSCHITZ CONDITION TO THAT IN THE EXISTING LITERATURE

In the existing literature, the function $f(x)$ is often assumed to be locally Lipschitz on any compact set $\mathcal{X} \subset \mathbb{R}$. This implies that for any compact set $\mathcal{X} = [a, b]$, there exists a constant $L_{\mathcal{X}} < \infty$ such that:

$$|f(x) - f(y)| \leq L_{\mathcal{X}}|x - y|, \quad \forall x, y \in \mathcal{X}.$$

For example, consider $f(x) = x^2$, where the derivative $f'(x) = 2x$. On $\mathcal{X} = [0, 100]$, the Lipschitz constant is:

$$L_{\mathcal{X}} = 2 \cdot \max_{x \in \mathcal{X}} |x| = 200.$$

This large constant makes the bound impractical in matching-based inference, where overly conservative bounds can restrict the formation of matched sets.

In contrast, our Derivative Control condition improves on the Lipschitz assumption by explicitly tying the slope of $f(x)$ to the size of the matched set. Specifically, it requires:

$$\sup_{x \in \mathcal{C}_t} |f'(x)| \cdot \text{radius}(\mathcal{C}_t) \leq M,$$

where:

- $\mathcal{C}_t$ is the matched set for a given t ,
- $\text{radius}(\mathcal{C}_t)$ is the diameter of the matched set in x -space,
- M is a universal constant independent of the matched set size.

This condition offers several practical advantages:

1. **Localized Control:** Instead of requiring a single large Lipschitz constant $L_{\mathcal{X}}$ over a wide range, our condition focuses on smaller, localized matched sets.
2. **Adaptive Bounds:** When the derivative $f'(x)$ is large, our condition naturally enforces smaller matched set radii to maintain practical bounds. For instance:

$$\text{If } f'(x) = 100 \text{ (as for } x = 50\text{), then } \text{radius}(\mathcal{C}_t) \leq \frac{M}{100}.$$

3. **Real-World Applicability:** In real-world matching problems, matched sets are typically small, and our condition aligns with this reality by providing sharper, more practical bounds than the overly conservative Lipschitz constant.

To summarize, while the Lipschitz assumption is valid on compact sets, the associated constants $L_{\mathcal{X}}$ can become impractically large for functions like $f(x) = x^2$ over wide intervals. By explicitly accounting for both the derivative and the size of matched sets, our condition provides a more precise and practical framework for matching-based inference.

C.6 COMPARISON WITH THEOREM 1 OF [125]

Our theorem differs from Theorem 1 of [125] in several key aspects. While both results address consistency in variance estimation under heteroskedasticity, the differences lie in the frameworks, assumptions, and proof strategies.

C.6.1 PARAMETRIC VS. NONPARAMETRIC FRAMEWORK

White’s Theorem 1 is based on a regression model $Y_i = X_i\beta_0 + \varepsilon_i$, where ε_i represents independent but non-identically distributed (i.n.i.d.) errors. The parametric form $X_i\beta_0$ is central, and β_0 is estimated via ordinary least squares (OLS). Heteroskedasticity arises through $\text{Var}(\varepsilon_i | X_i) = g(X_i)$, where $g(X_i)$ is a known (possibly parametric) function. In contrast, our theorem relies on a nonparametric matching estimator for treatment effects,

without assuming a parametric form for $f(X_i)$. Matching is governed by hyperparameters like the number of neighbors or the maximum matching radius, but these are not estimated from the data in the regression sense. Heteroskedasticity arises through $\sigma^2(X_i)$, where $\sigma^2(\cdot)$ is a uniformly continuous function.

C.6.2 WHITE'S SETUP: ESTIMATING $\text{Var}(\hat{\beta})$ VS. CLUSTER-BASED VARIANCE ESTIMATION

White's Theorem 1 focuses on the heteroskedasticity-consistent (HC) covariance matrix estimator for $\hat{\beta}$. It defines the matrix

$$\hat{V}_n = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 X_i' X_i, \quad \text{where} \quad \hat{\varepsilon}_i = Y_i - X_i \hat{\beta}.$$

White proves $\hat{V}_n \xrightarrow{\text{a.s.}} \bar{V}_n$, where $\bar{V}_n$ is the asymptotic covariance matrix of the regressors. Our theorem, on the other hand, defines cluster-level residual variance estimators s_t^2 for each treated unit $t \in \mathcal{T}$, given its matched controls $\mathcal{C}_t$. The overall variance estimator is

$$S^2 = \frac{1}{n_T} \sum_{t=1}^{n_T} s_t^2, \quad \text{where} \quad s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j \in \mathcal{C}_t} e_{tj}^2.$$

We prove $|S^2 - \frac{1}{n_T} \sum_{t=1}^{n_T} \sigma_t^2| \xrightarrow{\text{a.s.}} 0$, showing consistency for the average cluster variance.

C.6.3 HOMOSKEDASTICITY IN MATCHED CLUSTERS VS. GENERAL HETEROSKEDASTICITY

White's Theorem 1 allows general heteroskedasticity: $\text{Var}(\varepsilon_i | X_i) = g(X_i)$, where $g(\cdot)$ can vary arbitrarily across observations. Errors are independent but not identically distributed (i.n.i.d.). Our theorem also allows heteroskedasticity: $\sigma^2(X_i)$ varies with X_i . However, within each matched cluster $\{t\} \cup \mathcal{C}_t$, we assume $\sigma_j^2 \approx \sigma_t^2$ for $j \in \mathcal{C}_t$, based on a uniform continuity (or Lipschitz) assumption on $\sigma^2(\cdot)$.

C.6.4 PROOF STRATEGY AND KEY ASSUMPTIONS

White’s proof strategy relies on expanding $\hat{V}_n - \bar{V}_n$ and showing that

$$\hat{V}_n - \bar{V}_n = \frac{1}{n} \sum_{i=1}^n (\hat{\varepsilon}_i^2 X_i' X_i - E[\varepsilon_i^2 X_i' X_i]) \xrightarrow{\text{a.s.}} 0.$$

White uses assumptions on finite moments of ε_i and X_i (Assumptions 2–4 in [125]) and uniform integrability conditions. Our proof, in contrast, relies on showing that for matched clusters $\{t\} \cup \mathcal{C}_t$, the residual variance s_t^2 converges to the true variance σ_t^2 . We leverage uniform continuity of $\sigma^2(\cdot)$ to argue that $\sigma_j^2 \rightarrow \sigma_t^2$ as $\|X_{tj} - X_t\| \rightarrow 0$. We then apply a version of the Law of Large Numbers (LLN) for matched clusters.

C.6.5 SUMMARY OF DIFFERENCES

The key differences between White’s theorem and our theorem can be summarized as follows. First, White’s theorem is regression-based, while our theorem is matching-based. Second, White assumes a parametric model $Y_i = X_i \beta_0 + \varepsilon_i$, whereas our model assumes a nonparametric $f_1(X)$, $f_0(X)$. Third, White’s focus is on a robust covariance estimator for $\hat{\beta}$, while ours is on residual variance from matched clusters. Fourth, White allows fully general $g(X_i)$, whereas our clusters assume approximate homoskedasticity ($\sigma_j^2 \approx \sigma_t^2$). Finally, White’s framework has no matching hyperparameters, while ours depends on predefined criteria for matching (e.g., number of neighbors or radius).

C.7 FULL SIMULATION SETTINGS

C.7.1 SIMULATION SETTINGS FOR CALIPER MATCHING

DATA-GENERATING PROCESS

COVARIATES. We simulate:

- *Fifty treated units*, each drawn from multivariate normal distributions centered at $(0.25, 0.25)$ and $(0.75, 0.75)$, with covariance matrices

$$\begin{bmatrix} 0.1^2 & 0 \\ 0 & 0.1^2 \end{bmatrix}.$$

- *Two hundred and twenty-five control units*, each from multivariate normal distributions centered at $(0.25, 0.75)$ and $(0.75, 0.25)$, also with covariance matrices

$$\begin{bmatrix} 0.1^2 & 0 \\ 0 & 0.1^2 \end{bmatrix}.$$

- *One hundred additional control units* distributed uniformly on the unit square $[0, 1] \times [0, 1]$ to ensure an adequate overlap region.

OUTCOMES. For each unit with covariates (x_1, x_2) , we generate an outcome via:

$$Y = f_0(x_1, x_2) + Z \cdot \tau(x_1, x_2) + \epsilon,$$

where:

- f_0 is the density function for a multivariate normal distribution centered at $(0.5, 0.5)$ with covariance matrix

$$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

- $\tau(x_1, x_2) = 3x_1 + 3x_2$ is the true treatment effect.
- $\epsilon \sim N(0, 0.5^2)$ is homoskedastic measurement error.

OVERLAP VARIATION. We vary the degree of overlap by increasing the proportion of control units drawn uniformly from the unit square and reducing those centered at $(0.25, 0.75)$ and $(0.75, 0.25)$. For each overlap configuration, we run 500 Monte Carlo trials.

MATCHING AND ESTIMATION

Matching is performed using a synthetic-control-like optimization within local neighborhoods (calipers). We use an adaptive strategy for the caliper size: shrinking it to select no more than five units in dense regions, and expanding it to include at least one unit in sparse regions. With larger calipers, the optimization often selects only a few control units, which can be distant from the treated point. We refer readers to [26] for details.

C.7.2 SIMULATION SETTINGS FOR BOOTSTRAP VARIANCE ESTIMATION

DATA-GENERATING PROCESS

The Data Generating Process (DGP) is defined as:

$$\begin{aligned} & \{Y_i, Z_i, X_i\}_{i=1}^N, \\ & Y_i(1) = \tau + m(\|X_i\|) + \epsilon_i, \quad Y_i(0) = m(\|X_i\|) + \epsilon_i, \\ & Z_i = \mathbb{I}\{P(X_i) \geq v_i\}, \quad v_i \sim U[0, 1], \\ & P(X_i) = \gamma_1 + \gamma_2 \|X_i\|, \quad X_i = (X_{1i}, \dots, X_{ki})', \\ & X_{ji} = \xi_i |\zeta_{ji}| / \|\zeta_i\| \quad \text{for } j = 1, \dots, k, \\ & \xi_i \sim U[0, 1], \quad \zeta_i \sim N(0, I_k), \quad \epsilon_i \sim N(0, 0.2^2), \end{aligned}$$

The function $m(z)$ is defined as:

$$m(z) = 0.4 + 0.25 \sin(8z - 5) + 0.4 \exp(-16(4z - 2.5)^2)$$

Additional parameters include: Treatment Model Parameters: $\gamma_1 = 0.15$, $\gamma_2 = 0.7$; True Effect: $\tau = 0$; Number of Replicates: 100; Estimand: Average Treatment Effect on the Treated (ATT).

MATCHING METHOD AND POINT ESTIMATOR

We implement the paper’s matching procedure using 8 nearest neighbors (8-NN) with uniform weighting, assigning a weight of $1/8$ to each matched control unit.

Given a matching procedure, we define the matching estimator for the ATT as:

$$\hat{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \sum_{j \in \mathcal{C}_t} w_{jt} Y_j \right) \quad (\text{C.24})$$

where $\mathcal{C}_t$ is the set of matched controls, and w_{jt} represents the weight assigned to control unit j when matched to treated unit t .

DEBIASING METHOD

A debiasing model estimates the conditional mean function $\mu(z, x) = E[Y \mid Z = z, X = x]$. It is used to offset the bias to achieve valid inference (see Section 3.10.4 for discussion of the issue). The debiased estimator is defined as:

$$\tilde{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \left(Y_t - \hat{\mu}(0, X_t) - \sum_{j \in \mathcal{C}_t} w_{jt} (Y_j - \hat{\mu}(0, X_j)) \right) \quad (\text{C.25})$$

Additional implementation details include:

- **Model Choice:** Linear model
- **Training Data:** Control data only

³This is curve 6 in [93].

- **Cross-fitting:** Implemented by dividing the control data into two halves

VARIANCE ESTIMATORS

BOOTSTRAP VARIANCE ESTIMATOR.

- Step 1: Use data with $Z_i = 0$ to construct $\hat{\mu}(0, x) = \hat{E}[Y|Z = 0, X = x]$.
- Step 2: Construct debiased estimate for each treated unit $t \in \mathcal{T}$:

$$\tilde{\tau}_t = (Y_t - \hat{\mu}(0, X_t)) - \sum_{j \in \mathcal{C}_t} w_{jt} (Y_j - \hat{\mu}(0, X_j))$$

- Step 3: Construct the debiased estimator: $\tilde{\tau} = \frac{1}{n_t} \sum_{t \in \mathcal{T}} \tilde{\tau}_t$
- Step 4: Construct the debiased residuals $R_t = \tilde{\tau}_t - \tilde{\tau}$
- Step 5: Perform Wild bootstrap on $\{R_t\}$ with special sampling weights
- Step 6: Construct confidence interval from bootstrap distribution

POOLED VARIANCE ESTIMATOR.

- Step 1: Obtain the debiased estimator $\tilde{\tau}_t$
- Step 2: Estimate the variance using:

$$\hat{V} = S^2 \left(\frac{1}{n_T} + \frac{1}{\text{ESS}(\mathcal{C})} \right) \tag{C.26}$$

where S^2 is a pooled variance estimator across clusters of treated and matched controls

- Step 3: Construct the 95% confidence interval by $\tilde{\tau} \pm 1.96 * \sqrt{\hat{V}}$

C.7.3 KANG AND SCHAFER SIMULATION SETTINGS

DATA-GENERATING PROCESS

The Kang and Schafer data generating process is structured as follows:

1. Generate latent covariates $Z_1, Z_2, Z_3, Z_4 \sim \mathcal{N}(0, I_4)$ where I_4 is the 4×4 identity matrix.
2. Calculate propensity scores:

$$p(Z) = \frac{1}{1 + \exp(Z_1 - 0.5Z_2 + 0.25Z_3 + 0.1Z_4)} \quad (\text{C.27})$$

3. Generate treatment assignment as $T \sim \text{Bernoulli}(p(Z))$.
4. Define the outcome model:

$$f_0(Z) = \frac{210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4}{50} \quad (\text{C.28})$$

5. Generate potential outcomes with treatment effect τ :

$$Y(0) = f_0(Z) + \epsilon \quad (\text{C.29})$$

$$Y(1) = f_0(Z) + \tau + \epsilon \quad (\text{C.30})$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

6. The observed outcome is:

$$Y = Y(0)(1 - T) + Y(1)T = f_0(Z) + \tau \cdot T + \epsilon \quad (\text{C.31})$$

7. Transform the latent covariates Z to create the observed covariates X :

$$X_1 = \exp(Z_1/2) \tag{C.32}$$

$$X_2 = \frac{Z_2}{1 + \exp(Z_1)} + 10 \tag{C.33}$$

$$X_3 = \left(\frac{Z_1 \cdot Z_3}{25} + 0.6 \right)^3 \tag{C.34}$$

$$X_4 = (Z_2 + Z_4 + 20)^2 \tag{C.35}$$

In our simulations, we use $n = 500$ observations and set the true treatment effect $\tau = 0$.

References

- [1] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.
- [2] Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.
- [3] Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.
- [4] Alberto Abadie and Guido W Imbens. A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843, 2012.
- [5] Jaap H Abbring. The event-history approach to program evaluation. In *Modelling and Evaluating Treatment Effects in Econometrics*, pages 33–55. Emerald Group Publishing Limited, 2008.
- [6] Jaap H Abbring and Gerard J Van den Berg. Social experiments and instrumental variables with duration outcomes. 2005.
- [7] Alekh Agarwal, H Brendan McMahan, and Zheng Xu. An empirical evaluation of federated contextual bandit algorithms. *arXiv preprint arXiv:2303.10218*, 2023.
- [8] Emmanuel Akintoye, Doosup Shin, Paulino Alvarez, and Alexandros Briasoulis. State-level variation in waitlist mortality and transplant outcomes among patients listed for heart transplantation in the us from 2011 to 2016. *JAMA network open*, 3(12):e2028459–e2028459, 2020.
- [9] Hassan Khader Y Almathami, Khin Than Win, and Elena Vlahu-Gjorgievska. Empirical evidence of internal and external factors influencing users’ motivation toward teleconsultation use. *Telemedicine and e-Health*, 30(1):141–156, 2024.
- [10] Christopher SD Almond, Ravi R Thiagarajan, Gary E Piercey, Kimberlee Gauvreau, Elizabeth D Blume, Heather J Bastardi, Francis Fynn-Thompson, and TP Singh. Waiting list mortality among children listed for heart transplantation in the united states. *Circulation*, 119(5):717–727, 2009.

- [11] Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer US, New York, NY, 1993. ISBN 978-0-387-94519-4, 978-1-4612-4348-9. doi: 10.1007/978-1-4612-4348-9. URL <http://link.springer.com/10.1007/978-1-4612-4348-9>.
- [12] Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.
- [13] Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022.
- [14] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.
- [15] Baruch Awerbuch and Yossi Azar. Buy-at-bulk network design. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 542–547. IEEE, 1997.
- [16] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.
- [17] Samuel L Battalio, David E Conroy, Walter Dempsey, Peng Liao, Marianne Menicetas, Susan Murphy, Inbal Nahum-Shani, Tianchen Qian, Santosh Kumar, and Bonnie Spring. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.
- [18] Lauren Bell, Claire Garnett, Tianchen Qian, Olga Perski, Henry WW Potts, and Elizabeth Williamson. Notifications to improve engagement with an alcohol reduction app: Protocol for a micro-randomized trial. *JMIR research protocols*, 9(8):e18690, 2020. ISSN 1929-0748. doi: 10.2196/preprints.18690.
- [19] Niranjana Bidargaddi, Daniel Almirall, Susan Murphy, Inbal Nahum-Shani, Michael Kovalcik, Timothy Pituch, Haitham Maaieh, Victor Strecher, et al. To prompt or not to prompt? a microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. *JMIR mHealth and uHealth*, 6(11):e10123, 2018.
- [20] Aviroop Biswas, Paul I Oh, Guy E Faulkner, Ravi R Bajaj, Michael A Silver, Marc S Mitchell, and David A Alter. Sedentary time and its association with risk for disease incidence, mortality, and hospitalization in adults: a systematic review and meta-analysis. *Annals of internal medicine*, 162(2):123–132, 2015.

- [21] Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018.
- [22] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [23] Leo Breiman. *Classification and regression trees*. Routledge, 2017.
- [24] Leah Burns, Nadege Le Roux, Robert Kalesnik-Orszulak, Jennifer Christian, Mathias Hukkelhoven, Frank Rockhold, and John O’Donnell. Real-world evidence for regulatory decision-making: guidance from around the world. *Clinical Therapeutics*, 44(3):420–437, 2022.
- [25] Thomas M. Cascino, Jeffrey S. McCullough, Xiaoting Wu, Michael J. Pienta, James W. Stewart, Robert B. Hawkins, Alexander A. Brescia, Ashraf Abou El Ala, Min Zhang, Pierre-Emmanuel Noly, Jonathan W. Haft, Jennifer A. Cowger, Monica Colvin, Keith D. Aaronson, Francis D. Pagani, Donald S. Likosky, Michigan Congestive Heart Failure Investigators, Khalil M Nassar, Paul Tang, Grace Chung, Tessa Watt, Michael P Thompson, and Allison M Janda. Comparison of Evaluations for Heart Transplant Before Durable Left Ventricular Assist Device and Subsequent Receipt of Transplant at Transplant vs Nontransplant Centers. *JAMA Network Open*, 5(11):e2240646, November 2022. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.40646. URL <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2798133>.
- [26] Jonathan Che, Xiang Meng, and Luke Miratrix. Caliper synthetic matching: Generalized radius matching with local synthetic controls. *arXiv preprint arXiv:2411.05246*, 2024.
- [27] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
- [28] Swati Choudhry, Yunfei Wang, Susan W Denfield, Antonio G Cabrera, Jack F Price, Hari P Tunuguntla, Vikas R Dharnidharka, Charles E Canter, and William J Dreyer. A recipient risk prediction tool for short-term mortality after pediatric heart transplantation. *Transplantation*, 103(11):2434–2439, 2019.
- [29] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- [30] John Crowley and Marie Hu. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357):27–36, 1977.
- [31] Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 2023.

- [32] Ersilia M DeFilippis, Kiran K Khush, Maryjane A Farr, Amy Fiedler, Arman Kilic, and Michael M Givertz. Evolving characteristics of heart transplantation donors and recipients: Jacc focus seminar. *Journal of the American College of Cardiology*, 79(11):1108–1123, 2022.
- [33] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.
- [34] Walter Dempsey, Peng Liao, Pedja Klasnja, Inbal Nahum-Shani, and Susan A Murphy. Randomised trials for the fitbit generation. *Significance*, 12(6):20–23, 2015.
- [35] Walter Dempsey, Peng Liao, Santosh Kumar, and Susan A Murphy. The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments. *The annals of applied statistics*, 14(2):661, 2020.
- [36] Peter Diggle. *Analysis of longitudinal data*. Oxford university press, 2002.
- [37] Nicki A Dowling, Stephanie S Merkouris, George J Youssef, Dan I Lubman, Kathleen L Bagot, Chloe O Hawker, Hannah J Portogallo, Anna C Thomas, and Simone N Rodda. A gambling just-in-time adaptive intervention (gamblingless: In-the-moment): Protocol for a microrandomized trial. *JMIR Research Protocols*, 11(8):e38958, 2022.
- [38] Oliver Dukes, Mats J Stensrud, Riccardo Brioschi, and Aaron Hudson. Nonparametric tests of treatment effect homogeneity for policy-makers. *arXiv preprint arXiv:2410.00985*, 2024.
- [39] Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.
- [40] Roger W Evans, Diane L Manninen, Louis P Garrison, and Anthony M Maier. Donor availability as the primary determinant of the future of heart transplantation. *Jama*, 255(14):1892–1898, 1986.
- [41] Caroline A Figueroa, Adrian Aguilera, Bibhas Chakraborty, Arghavan Modiri, Jai Aggarwal, Nina Deliu, Urmimala Sarkar, Joseph Jay Williams, and Courtney R Lyles. Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. *Journal of the American Medical Informatics Association*, 28(6):1225–1234, 2021.
- [42] Lloyd D. Fisher and D. Y. Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health*, 20(1):145–157, May 1999. ISSN 0163-7525, 1545-2093. doi: 10.1146/annurev.publhealth.20.1.145. URL <https://www.annualreviews.org/doi/10.1146/annurev.publhealth.20.1.145>.

- [43] Thomas R. Fleming and David P. Harrington. *Counting processes and survival analysis*. Wiley series in probability and statistics. Wiley-Interscience, New York, 2005. ISBN 978-0-471-76988-0.
- [44] Dylan J. Foster and Vasilis Syrgkanis. Orthogonal Statistical Learning, June 2023. URL <http://arxiv.org/abs/1901.09036>. arXiv:1901.09036 [cs, econ, math, stat].
- [45] Zijun Gao and Trevor Hastie. Estimating heterogeneous treatment effects for general responses. *arXiv preprint arXiv:2103.04277*, 2021.
- [46] Jessica R Golbus, Walter Dempsey, Elizabeth A Jackson, Brahmajee K Nallamothu, and Predrag Klasnja. Microrandomized trial design for evaluating just-in-time adaptive interventions through mobile health technologies for cardiovascular disease. *Circulation: Cardiovascular Quality and Outcomes*, 14(2):e006760, 2021.
- [47] Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of econometrics*, 225(2):254–277, 2021.
- [48] Arun Gopalakrishnan and Young-Hoon Park. The impact of coupons on the visit-to-purchase funnel. *Marketing Science*, 40(1):48–61, 2021.
- [49] Government Digital Service. Micro-randomised trial. <https://www.gov.uk/guidance/micro-randomised-trial>, 2023. Accessed: 2023-06-24.
- [50] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [51] James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654, 1997.
- [52] James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.
- [53] Miguel A Hern’an. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010.
- [54] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.
- [55] Kosuke Imai and In Song Kim. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415, 2021.
- [56] Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. 2013.
- [57] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

- [58] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.
- [59] Nathan Kallus. Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.*, 21:62–1, 2020.
- [60] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. 2007.
- [61] Luke J. Keele, Eli Ben-Michael, Avi Feller, Rachel Kelz, and Luke Miratrix. Hospital quality risk standardization via approximate balancing weights. *The Annals of Applied Statistics*, 17(2), June 2023. ISSN 1932-6157. doi: 10.1214/22-AOAS1629. URL <https://projecteuclid.org/journals/annals-of-applied-statistics/volume-17/issue-2/Hospital-quality-risk-standardization-via-approximate-balancing-weights/10.1214/22-AOAS1629.full>.
- [62] Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects, August 2023. URL <http://arxiv.org/abs/2004.14497>. arXiv:2004.14497 [math, stat].
- [63] Kiran K. Khush, Rebecca Menza, John Nguyen, Jonathan G. Zaroff, and Benjamin A. Goldstein. Donor Predictors of Allograft Use and Recipient Outcomes After Heart Transplantation. *Circulation: Heart Failure*, 6(2):300–309, March 2013. ISSN 1941-3289, 1941-3297. doi: 10.1161/CIRCHEARTFAILURE.112.000165. URL <https://www.ahajournals.org/doi/10.1161/CIRCHEARTFAILURE.112.000165>.
- [64] Arman Kilic, Michael A Mathier, Gavin W Hickey, Ibrahim Sultan, Victor O Morell, Suresh R Mulukutla, and Mary E Keebler. Evolving trends in adult heart transplant with the 2018 heart allocation policy change. *JAMA cardiology*, 6(2):159–167, 2021.
- [65] Predrag Klasnja, Beverly L Harrison, Louis LeGrand, Anthony LaMarca, Jon Froehlich, and Scott E Hudson. Using wearable sensors and real time inference to understand human recall of routine activities. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 154–163, 2008.
- [66] Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S): 1220, 2015.
- [67] Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of heartsteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

- [68] John P. Klein and Melvin L. Moeschberger. *Survival Analysis*. Statistics for Biology and Health. Springer New York, New York, NY, 1997. ISBN 978-1-4757-2730-2 978-1-4757-2728-9. doi: 10.1007/978-1-4757-2728-9. URL <http://link.springer.com/10.1007/978-1-4757-2728-9>.
- [69] Jan-Niklas Kramer, Florian Künzler, Varun Mishra, Bastien Presset, David Kotz, Shawna Smith, Urte Scholz, Tobias Kowatsch, et al. Investigating intervention components and exploring states of receptivity for a smartphone app to promote physical activity: protocol of a microrandomized trial. *JMIR research protocols*, 8(1):e11540, 2019.
- [70] Emily B Kroska, Sydney Hoel, Amanda Victory, Susan A Murphy, Melvin G McInnis, Zachary N Stowe, and Amy Cochran. Optimizing an acceptance and commitment therapy microintervention via a mobile app with two cohorts: protocol for micro-randomized trials. *JMIR research protocols*, 9(9):e17086, 2020.
- [71] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, March 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1804597116. URL <https://pnas.org/doi/full/10.1073/pnas.1804597116>.
- [72] S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer New York, New York, NY, 2003. ISBN 978-1-4419-1848-2 978-1-4757-3803-2. doi: 10.1007/978-1-4757-3803-2. URL <http://link.springer.com/10.1007/978-1-4757-3803-2>.
- [73] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- [74] Jung-Ah Lee, Mona Choi, Sang A Lee, and Natalie Jiang. Effective behavioral intervention strategies using mobile health applications for chronic disease management: a systematic review. *BMC medical informatics and decision making*, 18:1–18, 2018.
- [75] Lingling Li and Tom Greene. Evaluating the effect of health care interventions using doubly robust estimation. *Statistics in Medicine*, 34(3):539–559, 2015.
- [76] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.
- [77] Peng Liao, Predrag Klasnja, Ambuj Tewari, and Susan A Murphy. Sample size calculations for micro-randomized trials in mhealth. *Statistics in medicine*, 35(12):1944–1971, 2016.
- [78] Peng Liao, Walter Dempsey, Hillol Sarker, Syed Monowar Hossain, Mustafa Al’Absi, Predrag Klasnja, and Susan Murphy. Just-in-time but not too much: determining treatment timing in mobile health. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(4):1–21, 2018.

- [79] Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.
- [80] Minghong Lin, Zhenhua Liu, Adam Wierman, and Lachlan LH Andrew. Online algorithms for geographical load balancing. In *2012 international green computing conference (IGCC)*, pages 1–10. IEEE, 2012.
- [81] Jana Lipkova, Tiffany Y Chen, Ming Y Lu, Richard J Chen, Maha Shady, Mane Williams, Jingwen Wang, Zahra Noor, Richard N Mitchell, Mehmet Turan, et al. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nature medicine*, 28(3):575–582, 2022.
- [82] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
- [83] Fang Liu and Demosthenes Panagiotakos. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, 22(1):287, 2022.
- [84] Xueqing Liu, Nina Deliu, and Bibhas Chakraborty. Microrandomized trials: Developing just-in-time adaptive interventions for better public health. *American Journal of Public Health*, 113(1):60–69, 2023.
- [85] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12017–12025, 2022.
- [86] Mayo Clinic. High blood pressure (hypertension) - diagnosis and treatment. <https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/diagnosis-treatment/drc-20373417>, 2021.
- [87] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.
- [88] Inbal Nahum-Shani, Lindsey N Potter, Cho Y Lam, Jamie Yap, Alexander Moreno, Rebecca Stoffel, Zhenke Wu, Neng Wan, Walter Dempsey, Santosh Kumar, et al. The mobile assistance for regulating smoking (mars) micro-randomized trial design protocol. *Contemporary clinical trials*, 110:106513, 2021.

- [89] Timothy NeCamp, Srijan Sen, Elena Frank, Maureen A Walton, Edward L Ionides, Yu Fang, Ambuj Tewari, Zhenke Wu, et al. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: micro-randomized trial. *Journal of medical Internet research*, 22(3):e15033, 2020.
- [90] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.
- [91] Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409, 2021.
- [92] John B O’Connell, Robert C Bourge, Maria Rosa Costanzo-Nordin, David J Driscoll, James P Morgan, EA Rose, and BF Uretsky. Cardiac transplantation: recipient selection, donor procurement, and medical follow-up. a statement for health professionals from the committee on cardiac transplantation of the council on clinical cardiology, american heart association. *Circulation*, 86(3):1061–1079, 1992.
- [93] Taisuke Otsu and Yoshiyasu Rai. Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732, 2017.
- [94] Neville Owen, Geneviève N Healy, Charles E Matthews, and David W Dunstan. Too much sitting: the population-health science of sedentary behavior. *Exercise and sport sciences reviews*, 38(3):105, 2010.
- [95] William F. Parker, Allen S. Anderson, Robert D. Gibbons, Edward R. Garrity, Lainie F. Ross, Elbert S. Huang, and Matthew M. Churpek. Association of Transplant Center With Survival Benefit Among Adults Undergoing Heart Transplant in the United States. *JAMA*, 322(18):1789, November 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.15686. URL <https://jamanetwork.com/journals/jama/fullarticle/2754791>.
- [96] Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050, 1994.
- [97] Richard F Potthoff, Max A Woodbury, and Kenneth G Manton. ”Equivalent Sample Size” and ”Equivalent Degrees of Freedom” Refinements for Inference Using Survey Weights Under Superpopulation Models. 2024.
- [98] Alyssa Power, Kurt R. Sweat, Alvin Roth, John C. Dykes, Beth Kaufman, Michael Ma, Sharon Chen, Seth A. Hollander, Elizabeth Profita, David N. Rosenthal, Lynsey Barkoff, Chiu-Yu Chen, Ryan R. Davies, and Christopher S. Almond. Contemporary Pediatric Heart Transplant Waitlist Mortality. *Journal of the American College of Cardiology*, 84(7):620–632, August 2024. ISSN 07351097. doi: 10.1016/j.jacc.2024.05.049. URL <https://linkinghub.elsevier.com/retrieve/pii/S0735109724075624>.

- [99] Tianchen Qian, Ashley E Walton, Linda M Collins, Predrag Klasnja, Stephanie T Lanza, Inbal Nahum-Shani, Mashfiqui Rabbi, Michael A Russell, Maureen A Walton, Hyesun Yoo, et al. The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychological methods*, 2022.
- [100] Mashfiqui Rabbi, Meredith Philyaw Kotov, Rebecca Cunningham, Erin E Bonar, Inbal Nahum-Shani, Predrag Klasnja, Maureen Walton, Susan Murphy, et al. Toward increasing engagement in substance use data collection: development of the substance abuse research assistant app and protocol for a microrandomized trial using adolescents and emerging adults. *JMIR research protocols*, 7(7):e9850, 2018.
- [101] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.
- [102] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 87(420):1013–1022, 1992.
- [103] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.
- [104] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- [105] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.
- [106] Donald B Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- [107] Azeem M Shaikh and Panos Toulis. Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 116(536):1835–1848, 2021.
- [108] Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353, 2005.
- [109] Donna Spruijt-Metz, Benjamin M Marlin, Misha Pavel, Daniel E Rivera, Eric Hekler, Steven De La Torre, Mohamed El Mistiri, Natalie M Golaszweski, Cynthia Li, Rebecca Braga De Braganca, et al. Advancing behavioral intervention and theory development for mobile health: the heartsteps ii protocol. *International journal of environmental research and public health*, 19(4):2267, 2022.
- [110] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.

- [111] Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics*, 225(2):175–199, 2021.
- [112] J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, 106(1):1–31, 2023. doi: 10.18637/jss.v106.i01. URL <https://www.jstatsoft.org/index.php/jss/article/view/v106i01>.
- [113] Terry Therneau, Beth Atkinson, Brian Ripley, and Maintainer Brian Ripley. Package ‘rpart’. Available online: [cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf](https://cran.r-project.org/web/packages/rpart/rpart.pdf) (accessed on 20 April 2016), 2015.
- [114] David Thompson. Replication of randomized, controlled trials using real-world data: what could go wrong?, 2021.
- [115] Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255, 2022.
- [116] Elbert P Trulock, Jason D Christie, Leah B Edwards, Mark M Boucek, Paul Aurora, David O Taylor, Fabienne Dobbels, Axel O Rahmel, Berkeley M Keck, and Marshall I Hertz. Registry of the international society for heart and lung transplantation: twenty-fourth official adult lung and heart–lung transplantation report—2007. *The Journal of heart and lung transplantation*, 26(8):782–795, 2007.
- [117] Catrine Tudor-Locke and David R Bassett. How many steps/day are enough? preliminary pedometer indices for public health. *Sports medicine*, 34:1–8, 2004.
- [118] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [119] Stijn Vansteelandt and Marshall Joffe. Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science*, 29(4), November 2014. ISSN 0883-4237. doi: 10.1214/14-STS493. URL <https://projecteuclid.org/journals/statistical-science/volume-29/issue-4/Structural-Nested-Models-and-G-estimation--The-Partially-Realized/10.1214/14-STS493.full>.
- [120] Shresth Verma, Aditya Mate, Kai Wang, Neha Madhiwalla, Aparna Hegde, Aparna Taneja, and Milind Tambe. Restless multi-armed bandits for maternal and child health: Results from decision-focused learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1312–1320, 2023.
- [121] Joannes Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer, 2005.

- [122] Jitao Wang, Yu Fang, Elena Frank, Maureen A Walton, Margit Burmeister, Ambuj Tewari, Walter Dempsey, Timothy NeCamp, Srijan Sen, and Zhenke Wu. Effectiveness of gamified team competition as mhealth intervention for medical interns: a cluster micro-randomized trial. *npj Digital Medicine*, 6(1):4, 2023.
- [123] Yuyao Wang, Andrew Ying, and Ronghui Xu. Learning treatment effects under covariate dependent left truncation and right censoring, November 2024. URL <http://arxiv.org/abs/2411.18879>. arXiv:2411.18879 [stat].
- [124] Lee-Jen Wei. The adaptive biased coin design for sequential experiments. *The Annals of Statistics*, 6(1):92–100, 1978.
- [125] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.
- [126] Shenbo Xu, Raluca Cobzaru, Stan N. Finkelstein, Roy E. Welsch, Kenney Ng, and Zach Shahn. Estimating Heterogeneous Treatment Effects on Survival Outcomes Using Counterfactual Censoring Unbiased Transformations, September 2024. URL <http://arxiv.org/abs/2401.11263>. arXiv:2401.11263 [stat].
- [127] Yizhe Xu, Nikolaos Ignatiadis, Erik Sverdrup, Scott Fleming, Stefan Wager, and Nigam Shah. Treatment Heterogeneity for Survival Outcomes, September 2022. URL <http://arxiv.org/abs/2207.07758>. arXiv:2207.07758 [stat].
- [128] Junzhe Zhang and Elias Bareinboim. Online reinforcement learning for mixed policy scopes. *Advances in Neural Information Processing Systems*, 35:3191–3202, 2022.
- [129] Kelly W Zhang, Lucas Janson, and Susan A Murphy. Statistical inference after adaptive sampling for longitudinal data. *arXiv preprint arXiv:2202.07098*, 2022.
- [130] Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Mining electronic health records using real-world evidence for clinical research. *IEEE Intelligent Systems*, 32(6):50–56, 2017.
- [131] Jie Zhu and Blanca Gallego. Causal inference for observational longitudinal studies using deep survival models. *Journal of Biomedical Informatics*, 131:104119, 2022.
- [132] Yuanjia Zhu, Bharathi Lingala, Michael Baiocchi, Veronica Toro Arana, Kiah M Williams, Yasuhiro Shudo, Philip E Oyer, and Y Joseph Woo. The stanford experience of heart transplantation over five decades. *European Heart Journal*, 42(48):4934–4943, 2021.
- [133] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

- [134] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.