Covariate Adjustment in Modern Causal Design and Analysis

by

Lauren Diana Liao

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Samuel D. Pimentel, Co-chair
Professor Alejandro Schuler, Co-chair
Professor Alan E. Hubbard
Doctor Yeyi Zhu

Summer 2024

Covariate Adjustment in Modern Causal Design and Analysis

Abstract

Covariate Adjustment in Modern Causal Design and Analysis

by

Lauren Diana Liao

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Samuel D. Pimentel, Co-chair

Professor Alejandro Schuler, Co-chair

Clinicians, policymakers, psychologists, and economists often ask causal questions: does this treatment or intervention cause the observed differences in outcome? If researchers can observe both outcomes under treatment (intervention or exposure) and control for each unit, then researchers can directly measure the causal effect. However, the inability to observe both outcomes forms the fundamental problem of causal inference. Answers to these questions require formal frameworks to interpret statistical quantity with causality. While the gold standard of causality stems from randomized controlled trials or experiments due to treatment assignment randomization, observational studies are increasingly popular for researchers to discover existing phenomena and surveillance to provide real-world evidence. Covariates, measured alongside treatment and outcome, are used in modern causal inference to improve analyses in observational and experimental studies. This dissertation proposes methods to thoughtfully identify and adjust for covariates in observational study design, risk factor analysis, and randomized trial analysis. In Chapter 2, we advocate for a new visualization tool, the joint variable importance plot, to help researchers prioritize confounders for adjustment in observational study design. In Chapter 3, we present variable importance from prediction and as-if causal perspective to evaluate mortality risk factors. Lastly, we encourage practitioners to adopt prognostic covariate adjustment with efficient estimators when analyzing small randomized trials in Chapter 4.

To all my loved ones who believed in me.

Contents

List of Figures

List of Tables

Acknowledgments

My loved ones, friends, colleagues, and mentors have been integral to this journey, for whom I am deeply grateful for the support I have received. To Jon, your unwavering love and encouragement have been a constant source of strength. To Sharon, Allison, Kathryn, Raunak, Cory, and Evelyn, your friendship supported me through challenging times. To the Zhu group at Kaiser Permanente Northern California, Division of Research, especially Amanda and Emily, thank you for helping me ground my research in real-world contexts. To my friends in the biostatistics and statistics programs (casual causal group), I am thankful for learning and growing together in this journey. To Pratik Sachdeva, your mentorship has helped me become a better communicator. To Alan Hubbard and Yeyi Zhu, you inspire me to become a better collaborative researcher, and I am incredibly fortunate to have worked with you. To my advisors, Sam Pimentel and Alejandro Schuler, I am immensely grateful for your continuous generosity, kindness, and guidance in shaping my research.

Chapter 1

Introduction

While statisticians traditionally emphasize associations, causality is often the true question of interest. Causal inference focuses on measuring the impact of treatment or intervention on an outcome, which can be used to inform economical, political, and clinical decisions. Recent technological advancement lead to the current era of data abundance, such that many covariates measured prior to treatment became available. Adjusting for these covariates is important because if the covariates are confounders – meaning they are related to both treatment assignment and outcome, then they potentially can explain away or mask the true causal effect. Covariate adjustment can also decrease the estimated variance, thus providing a more efficient inference. A unique set of challenges arise when considering the interpretation of important covariates in these high-dimensional problems. Especially for methods that provide valid inference without imposing strong parametric assumptions, careful interpretation of variable importance on these covariates are necessary for better applications on real-world data.

The three contributions from the dissertation aim to improve causal design and analysis through covariate adjustment to identify important variables, clarify interpretation, and reduce uncertainty. This thesis builds on elements of machine learning, variable importance in prediction and causal interpretation, and leveraging external, historical (prior collected) data to achieve these goals.

In Chapter 2 we introduce a novel visualization, namely the joint variable importance plot, to help researchers prioritize covariates in their observational study design. Although randomized controlled trials (RCTs) or experiments are the gold standard to evaluate a causal effect due to the randomized treatment assignment, observational studies provide insights to impact of existing phenomena. Observational study design focuses on establishing balance, or similar covariate profiles between treated and control groups, prior to analysis. Note that in an observational context, treated group requires careful definition of "treatment" since there is no active assignment; alternatively, treated versus controls groups can be considered as with or without exposure, respectively. We advocate for prioritizing variables that may be strong confounders as related to both treatment imbalance and outcome importance (informed by external pilot data) for adjustment in subsequent analysis

stage. This visualization can be integrated to many modern design approaches in matching, weighting, and regression[135]. The contents of this chapter have been published in Liao et al.[73].

In Chapter 3 we analyze mortality risk factors of those infected by SARS-CoV-2 in prediction and as-if causal approaches. We apply the super learner[88,120] to evaluate risk factors important to prediction, and targeted maximum likelihood estimation[89,118,119] to further investigate the impact of pre-existing conditions, such as diabetes, hypertension, and obesity. We not only advocate for using machine learning to model non-linear relationship in mortality risk estimation but also highlight the importance of different variables throughout the three phases of COVID-19 pandemic: phase 1 is from March 1st, 2020, to October 31st, 2020, phase 2 is from November 1st, 2020, to March 31st, 2021, and phase 3 is from April 1st, 2021, to November 3rd, 2021. This chapter motivates researchers to use machine learning to identify those at risk and judiciously interpret risk factors and their importance. The contents of this chapter have been published in Liao et al.[72].

Lastly, Chapter 4 we propose prognostic covariate adjustment with efficient estimators to increase efficiency in small, finite RCTs. Prognostic score formalized by Hansen[49] characterizes predictions using trial covariates from an outcome model extracted from an external data set. This chapter extends prognostic covariate adjustment with linear estimators[104] to adjustment with efficient estimators to include external, historical data and increase efficiency without inflating type I error. We demonstrate empirical and theoretical improvement using this method to reduce finite sample uncertainty.

Chapter 2

Prioritizing Variables for Observational Study Design Using the Joint Variable Importance Plot

This chapter introduces the joint variable importance plot that visualizes natural prioritization of all potential confounders for adjustment in observational study design. The coauthors are Yeyi Zhu, Amanda L. Ngo, Rana F. Chehab, and Samuel D. Pimentel[73]. We gratefully acknowledge support from *Hellman Fellowship, National Science Foundation 2142146 and DGE 2146752, National Institute of Diabetes and Digestive and Kidney Diseases K01DK120807, National Heart, Lung, and Blood Institute R01HL157666, and Kaiser Permanente Northern California Community Benefits Program RNG209492*. The authors thank David Bruns-Smith, Avi Feller, Erin Hartman, Melody Y. Huang, Yaxuan Huang, Sizhu Lu, Arisa Sadeghpour, Andy Shen, and Arnout van Delden for valuable comments.

2.1 Introduction

Researchers often seek to evaluate treatments to understand whether they are beneficial. In observational (non-randomized) studies, treatments may be confounded, or associated with other baseline variables so that it is unclear whether to attribute group outcome differences to treatment or baseline dissimilarity. To reliably estimate an effect, researchers must adjust for these variables, typically either by modeling their impact on study outcomes or by creating new comparison groups that eliminate baseline differences or imbalances, for example by matching or weighting.

One crucial decision is deciding which variables are most important for adjustment. While creating comparison groups with perfect balance on the joint distribution of all baseline variables, or conditioning appropriately on this joint distribution in an outcome model, is sufficient to remove observed sources of confounding, this is impossible in datasets with a large number of measured variables. Attempting to adjust for too many variables can lead

to undesirable designs, such as heavily overfitted models, matches with too few subjects to be useful[134], or weighting designs with high-variance weights that hurt precision[79]. Many modern causal inference methods are designed with variable prioritization in mind and incorporate substantive or data-driven knowledge about which variables are likely to matter most. These include regularization procedures for outcome regression[8], balance tolerances for weighting[13], and covariate distances or balancing constraints for matching[14,86,113]. However, there is a need for better data-driven diagnostic tools to guide researcher choices about prioritization.

Researchers may be tempted to prioritize variables based on standard balance diagnostics, including tables of standardized mean differences (SMD) for each variable or Love plots[3,43,50,97,114]. These diagnostics are useful for highlighting variables with large imbalances between treated and control groups. However, prioritizing variables according to their imbalance ignores important information about the role of each variable in the outcome model. Variables strongly related to treatment but unrelated to outcomes are *not* confounders. In contrast, if variables are strongly associated with the outcome but with only moderate imbalance, they may be *strong* confounders. When choosing which baseline variables to prioritize for adjustment, focusing solely on the treatment imbalance can risk ignoring variables that should take precedence due to their outcome importance.

The joint importance of covariate-treatment and covariate-outcome relationships is a general principle in observational causal inference, not specific to a particular framework or set of identification assumptions. For example, outcome regression approaches typically do not make assumptions about the treatment-covariate relationship, but these relationships influence treatment effect estimation (see Section 2.2). Similarly, matching and weighting approaches are typically motivated by models of the treatment variable in covariates, but similarity of outcomes within matched pairs or across weighted groups affects residual bias[12,100]. Another reason outcome-covariate relationships matter is their influence on sensitivity to unmeasured bias. In both matching and weighting, increasing homogeneity of the outcomes via better control for prognostic covariates increases robustness to worst-case confounding as measured by design sensitivity[62,92]. Unfortunately design sensitivity is understudied in observational study design, and diagnostic tools to improve it are badly needed.

To meet these needs we propose selecting high-priority variables for adjustment using the joint treatment-outcome variable importance plot (jointVIP). JointVIP represents each variable in two dimensions: one describing treatment model importance as measured by the SMD, and one describing outcome-model importance, measured by outcome correlation among controls from a pilot sample (chosen disjointly from the analysis sample to ensure the integrity of the analysis). In addition, under a set of simple working models, the bias incurred by ignoring each variable can be derived separately and represented on the plot using unadjusted bias curves, enhancing opportunities for variable comparisons. We show an example comparison between the traditional Love plot and jointVIP with a subset of the baseline variables from the case study (absolute measures shown in Figure 2.1 and signed measures shown in Appendix A.1).

We illustrate jointVIP in detail in a case study of drug safety for diabetes medication in

pregnant individuals. Specifically, we use a matched design with refined covariate balance constraints, which require a prioritized list of variables to be specified for balancing, and jointVIP provides a principled way to choose this. However, we argue that jointVIP's value is not specific to a given estimation strategy or set of identification assumptions.

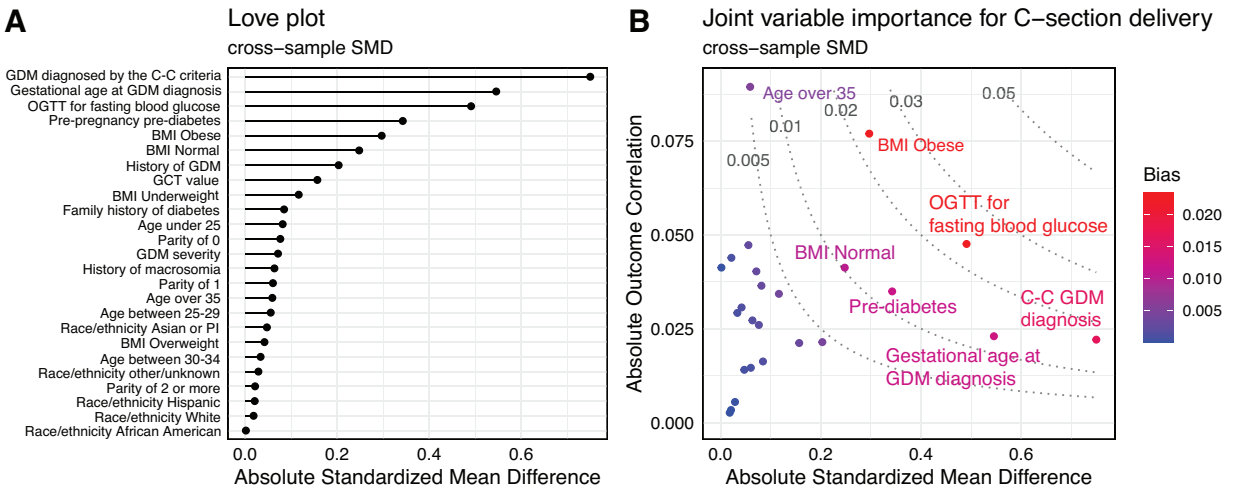


Figure 2.1: Comparison between the Love plot and the joint variable importance plot (jointVIP). Note that some variables (body mass index in the obese category and oral glucose tolerance test for fasting blood glucose used at gestational diabetes diagnosis) take on much more prominent positions in jointVIP than in the Love plot, which only displays standardized mean difference values.

2.2 Method

Joint variable importance plot construction

The high-level purpose of the jointVIP is to illustrate two different dimensions of a variable's possible role as a confounder – its imbalance, or association with the treatment variable, and its association with the outcome – on two axes, with each variable plotted as a single point. We now discuss the specific measures of variable importance on each axis.

For treatment model importance, described by the x-axis, we use SMDs, or differences between the treated mean and the control mean divided by an estimate of the variable's standard deviation. Many different standard deviation estimates have been proposed leading to slightly different SMD definitions; we focus on a version denoted as the "cross-sample" SMD, which uses the sample standard deviation of the variable in question computed in the pilot (control) sample. For more motivation and discussion of the cross-sample SMD, see Section 2.2. The variant we propose is similar to an effect size estimator from Glass [40],

which standardizes the mean difference by dividing by the standard deviation from the control group[52]. SMDs allow intuitive comparisons across variables with very different scales, including both binary and continuous variables. They are widely used to assess imbalance and are commonly reported in balance tables or Love plots. Thus, using SMD on the x-axis allows jointVIP to preserve all information typically contained in the Love plot while adding new insights.

For outcome model importance, represented on the y-axis, we compute the sample Pearson correlation between each variable and the outcome among controls. Sample correlation is a familiar, bounded quantity and makes sense for relationships not only between two continuous variables but also between two binary variables (phi coefficient), and between binary and continuous variables (point biserial correlation)[83]. The outcome relationship is calculated only among controls to avoid having to model treatment effects.

It is vital that the outcome correlations be computed in a pilot sample separate from the data used for the ultimate outcome analysis. Using controls from the analysis sample for computing outcome correlations can bias treatment effect estimates. For example, suppose treated and control samples exhibit imbalance on several continuous background variables (with treated individuals taking larger values), but the study outcome is independent of all these variables in the population. If we compute sample outcome correlations in the analysis control sample and form matched pairs based solely on the variable with the largest such (positive) sample correlation, we essentially match on the variable with the largest spurious correlation (with the random outcome noise in the current sample). Because of the imbalance, the matching algorithm will systematically select controls with large values for the spuriously correlated variable. Hence, the result will have large positive outcome errors that introduce positive bias into the average outcome for the matched controls. For related examples see Hansen[49] and Abadie et al.[1].

To construct a pilot sample, one may select a small (10-20%) portion of the control sample at random from the full control group. To ensure the pilot sample is drawn from the portion of the control space most relevant for the observational study, Aikens et al.[5] instead suggest conducting an initial round of matching on a standard Mahalanobis distance to pair each treated subject to two controls, then selecting one control from each set at random to construct the pilot sample. Alternatively, external data separate from the analysis of interest may be used as a pilot sample.

Addition of unadjusted bias curves for variable comparison

Comparing the relative importance of two distant points on the jointVIP, one with a high outcome correlation and low SMD, and the other with a high SMD and low outcome correlation, can be difficult. A natural answer lies in the relative sizes of the biases contributed by ignoring each variable, since our ultimate goal is to avoid biases in treatment effect estimation. We consider each baseline variable and evaluate the bias incurred by omitting this potential confounder under a simple one-variable model. Inspired by Cinelli and Hazlett[24] and Soriano et al.[110], we plot these bias estimates as curves on the jointVIP.

For any baseline variable X_j with $j \in 1, ..., J)$, consider the sample least-squares fit of outcome Y on baseline variable X_j and binary treatment Z:

$$Y = Z\tau_0 + X_j\beta_j + \hat{\epsilon} \tag{2.1}$$

Here $\hat{\epsilon}$ is a residual. In addition, consider two related sample regressions:

$$Y = Z\tau + \hat{e} \tag{2.2}$$

$$X_j = Z\Delta_j + \hat{u} \tag{2.3}$$

Following Cochran's formula[27], we may use (2.3) to rewrite (2.1) and obtain a new representation for (2.2):

$$Y = X_j\beta_j + Z\tau_0 + \hat{\epsilon} = (Z\Delta_j + \hat{u})\beta_j + Z\tau_0 + \hat{\epsilon} = Z(\Delta_j\beta_j + \tau_0) + (\hat{u}\beta_j + \hat{\epsilon}) \tag{2.4}$$

Since the new error term $(\hat{u}\beta_j + \hat{\epsilon})$ is orthogonal to Z by the construction of residuals $\hat{u}$ and $\hat{\epsilon}$, we have $\tau = (\Delta_j\beta_j + \tau_0)$ and $\hat{e} = \hat{u}\beta_j + \hat{\epsilon}$. Note that until now we have made no model assumptions, merely fit regressions using sample quantities; however, if we add a working assumption that triples (X, Y, Z) are sampled independently from an infinite population, with model (2.1) correctly specified (i.e. that $E(Y|X_j, Z) = \beta_j^{pop}X_j + \tau^{pop}Z$ for some parameters β_j^{pop} and τ^{pop}), then the difference

$$\tau - \tau_0 = \Delta_j\beta_j \tag{2.5}$$

is an estimate of the large-sample omitted variable bias (OVB) incurred by estimating treatment effects via regression on Z alone, ignoring X_j.

Importantly for our purposes, the OVB can be rewritten in terms of sample correlation between X_j and Y and a SMD with normalization by the control standard deviation. The key is that when equation (2.1) is fit on controls alone (as it will be in our pilot-sample approach), both (2.1) and (2.3) are simple regressions. We rewrite the corresponding simple regression equations using familiar simple regression formulae. $S_{Y_{pilot}}$ and $S_{X_{j,pilot}}$ denote the standard deviation of the pilot sample for outcome and the standard deviation of the confounder in question respectively. We include the *pilot* and *analysis* notations for clarity.

$$\beta_j = r_{X_{j,pilot},Y_{pilot}}\frac{S_{Y_{pilot}}}{S_{X_{j,pilot}}} \tag{2.6}$$

$$\Delta_j = \bar{X}_{j1,analysis} - \bar{X}_{j0,analysis} \tag{2.7}$$

Using (2.3), we obtain (2.7), where $\bar{X}_{j1,analysis}$ and $\bar{X}_{j0,analysis}$ denote variable j's sample means among treated subjects and controls, respectively, in the analysis sample. Substituting into expression (2.5) and rearranging, we obtain:

$$\frac{\Delta_j\beta_j}{S_{Y_{pilot}}} = r_{X_{j,pilot},Y_{pilot}}\frac{(\bar{X}_{j1,analysis} - \bar{X}_{j0,analysis})}{S_{X_{j,pilot}}} \tag{2.8}$$

The left-hand side is a conveniently normalized version of the OVB that is invariant to rescalings of the outcome, and the right-hand side is a product between a sample correlation computed in the pilot sample and a standardized difference defined as follows:

$$\text{cross-sample SMD} = \frac{\bar{X}_{j1,analysis} - \bar{X}_{j0,analysis}}{S_{X_{j,pilot}}} \tag{2.9}$$

The SMD calculates the difference between treated and control groups from the analysis sample and is standardized by the standard deviation from the pilot sample. Hence, we define this SMD as "cross-sample SMD".

Since the standardized OVB is a product of two terms, level sets for bias take the form of hyperbolic curves on the jointVIP to demarcate equivalent levels of confounding under the crude one-confounder models. In addition, a measure of bias may be computed for any individual variable using its respective SMD and outcome correlation, and color-coding based on these quantities is used for plotting points. We refer to the marginal bias measure as "unadjusted bias" to distinguish from the typical multivariate OVB models.

Bias in a finite population framework

The bias analysis of Section 2.2 assumes covariates, treatments, and outcomes are sampled jointly from an infinite population. Although the case study in Section 2.3 instead uses a finite population framework, this analysis still turns out be relevant. Given K matched pairs (with the treated unit indexed $k1$ in each pair k and the control unit indexed $k2$), that unobserved confounding is absent, and that $Y_{ki}(1) - Y_{ki}(0) = \tau$ for all k, i. The bias of a matched difference-in-means estimator for τ, viewing only Z as a random variable and holding potential outcomes $Y(1), Y(0)$ and covariates X fixed, can be written as follows:

$$\frac{1}{K}\sum_{k=1}^{K}[Y_{k1}(0) - Y_{k2}(0)](p_{k1} - p_{k2}) \tag{2.10}$$

where $p_{ki} = \frac{\lambda_{ki}/(1-\lambda_{ki})}{\lambda_{k1}/(1-\lambda_{k1})+\lambda_{k2}/(1-\lambda_{k2})}$ with λ_{ki} representing the propensity score for unit ki; for derivations see Sales et al.[100], §4 and Huang and Pimentel[60]. This formula suggests that attention to covariate-outcome relationships can improve estimation and inference via reduction in the magnitude of the $Y_{k1}(1) - Y_{k2}(0)$ terms. In principle it would vanish if matching were exact on the propensity score, but in practice this is implausible[45,85]. Additionally, if we consider the expected behavior of this bias when potential outcomes are drawn from a model and covariate X is ignored, we arrive at an approximate bound that is a rescaled version of unadjusted bias (see the Appendix A.2 for full derivation). Under similar assumptions, Rosenbaum[92] shows that reducing the variance of the $Y_{k1}(0) - Y_{k2}(0)$ terms reduces sensitivity to unmeasured bias, even when propensity score matching is exact. In summary, although in Section 2.2 we did not explicitly motivate the unadjusted bias curves in the context of biases

incurred under matched designs nor explicitly invoke the finite-sample framework typically used to analyze such designs, the tools developed in Section 2.2 retain useful interpretations from the perspective of matched analysis. We anticipate similar benefits for other causal inference strategies.

Using jointVIP to guide study design

Once the jointVIP has been created, researchers can select variables with large potential bias contributions (as measured by the unadjusted bias curves) for adjustment or otherwise leverage its information to choose tuning parameters. In a matched study, selected variables might be used to create a Mahalanobis distance[48], or their marginal imbalance could be restricted via fine or refined balance constraints[86,128] as in our case study in Section 2.3. In a study using stable balancing weights inverse values of the outcome correlations plotted on the y-axis of the jointVIP could be used as balance tolerances[133]. In outcome regression settings where the data is too high-dimensional to allow inclusion of all covariates, variables highlighted by jointVIP could be chosen as regressors. For matched and weighted studies, a post-design version of the jointVIP can also be created using new SMDs computed on the matched or weighted data. This can suggest further refinements of the original matching or weighting specification, or whether residual bias is large enough to require additional regression adjustment after matching and weighting and which variables should be included in such an adjustment model[91]. Table 2.1 summarizes the process of creating and applying jointVIP for practitioners, and a simulation study in Appendix A.3 empirically demonstrates the value of this process for bias reduction.

A natural question is how or whether to combine the process just described with the balance testing approach to study design proposed by Hansen and Bowers[50] for matched or stratified observational studies. Here the design is improved iteratively until an omnibus test using all measured covariates fails to reject the hypothesis that treatment is distributed uniformly within strata. While the jointVIP framework offers important new information by leveraging outcome-covariate relationships ignored by balance tests, balance tests offer a clearer ideal benchmark for success in the form of a hypothetical study randomized within strata, and a single condition to check incorporating all covariates. A researcher might proceed by requiring the final stratified design both to pass a balance test and to minimize potential bias as computed under jointVIP to enjoy the benefits of both frameworks. If this proves impossible, a researcher might instead use jointVIP to select a priority subset of covariates with highest outcome correlation, and search for a design for which the tests of Hansen and Bowers[50] fail to detect differences with respect to these variables. For an interesting related proposal to use prognostic information to construct a test statistic for balance testing, see Bicalho et al.[16].

JointVIP can also draw attention to variables with high treatment-model importance but negligible outcome-model importance, sometimes referred to as instrumental variables or prods[87]. Even when all variables could be used for adjustment, it is wise to exclude to such variables since they can inflate unmeasured confounding bias[18,33]. JointVIP enables

1. choose pilot sample	define pilot sample either as hold-out set from main analysis sample or from external historical data
2. create the jointVIP	fit outcome correlations from the pilot sample and compute SMD from the analysis sample
3. identify potential confounders	prioritize variables in top right region of the plot and use bias curves to make fine distinctions
4. adjust for confounders	create balance constraints (matching or weighting), a covariate distance (matching), a regressor matrix (outcome regression), etc. using chosen variables.
5. (optional) plot post-jointVIP	for matching and weighting re-plot with post-design SMD
repeat steps 3-5	repeat as desired

Table 2.1: Suggested procedure for use of the joint variable importance plot. As discussed in Section 2.2, the pilot sample typically consists of controls only. See Section 2.2 for further details on practical use of joint variable importance plot.

either excluding such variables or (if it is not entirely clear whether a variable should be excluded) constructing multiple control groups that adjust for these variables differently[87].

Some caution should be exercised when using and interpreting jointVIP. Outcome correlations can change substantially when variables are transformed; outliers may also skew the means of either treatment or control groups and hence the standardized mean differences. Blindly using all variables above a bias cutoff may also be suboptimal. For example, if two variables are near-perfectly collinear, both would be highlighted as priorities in jointVIP, but adjusting for one may be sufficient to remove bias. Finally, baseline variables that are absent or rare in the pilot sample may not be well-represented in the plot.

2.3 Case Study

Glyburide as a treatment for gestational diabetes

Due to improved ease of use and lower cost, oral antidiabetic medications, such as glyburide, are often prescribed compared to the recommended insulin therapy as treatment for gestational diabetes[19]. The safety of glyburide, however, remains contentious due to potential transfer to the fetus through the placenta[7]. The question remains: does glyburide increase the risk of adverse perinatal outcomes in real-world settings? We investigate glyburide's impact on C-section delivery compared to medical nutritional therapy, the universal first-line therapy in a large, population-based cohort.

The study population consists of Kaiser Permanente Northern California (KPNC) members. Individuals who are diagnosed with GDM receive medical nutritional therapy (MNT)

as the universal first line of therapy. Pharmacologic treatment, including oral antidiabetic medications (glyburide, metformin, or other) and/or insulin, is prescribed in addition to MNT if glycemic control goals are not met. Individuals with GDM who received MNT alone constituted our control group while those who additional received glyburide as the only pharmacologic therapy constituted our treatment group. There are 54 common variables between the 2007-2010 data (pilot sample) and 2011-2021 data (analysis sample), including indicators of missing data as variables. Table 2.2 summarizes selected baseline variables (see Appendix A.4 for the full data summary). Missing values were imputed separately for each year using random forest[112]. Details about the pattern of missing values and the imputation procedure are reported in Appendix A.5. Our use of KPNC data for this study is approved by the KPNC Institutional Review Board, which waived the requirement for informed consent from participants.

Design

Variable selection using jointVIP

JointVIP is constructed using the `jointVIP` package in R; for a brief software tutorial see Liao and Pimentel[71]. To ensure particularly stringent control of the propensity score, we impose a caliper equal to 0.2 standard deviations of the fitted propensity score values in the entire sample. Using a caliper on the propensity score is a natural choice because our approach to inference relies on similar propensity scores within matched pairs[85]. We match exactly on year to address substantive concerns about potential for temporal shifts in the standard of care in the absence of reliable outcome correlations.

We address potential bias from additional variables by imposing a series of refined balance constraints tailored to the outcome. Refined covariate balance enables users to specify top-priority variables and their interactions to be balanced as though they were the only variables in the study, with lower-priority variables receiving further attention as possible[86]. While this framework offers substantial flexibility to the researcher, it relies on strong substantive knowledge to specify the balance tiers in a reasonable manner. Frequently it is not immediately clear how to organize a group of baseline variables into balance tiers in a principled way. JointVIP offers a data-driven approach in settings where ambiguity remains even after accounting for substantive knowledge. We specify tiers of variables for refined covariate balance by identifying sets of variables with high importance. Since the prognostic score (fit in the pilot sample using LASSO regression) ranks among the variables contributing the largest unadjusted bias, we include quintiles of the prognostic score in the first balance tier. We include all variables contributing unadjusted bias greater than or equal to 0.010 with variables in subsequent tiers, with those contributing larger amounts of bias in higher tiers. Table 2.3 summarizes the chosen balance tiers for the design. Specific potential bias values can be found in Appendix A.6 column *Pre-matched bias*. In addition, we discretized the continuous variable for gestational age at GDM diagnosis for compatibility with refined covariate balance algorithm.

Table 2.2: Summary of selected baseline variables for pregnant individuals with gestational diabetes.

		2007-2010 Control n = 7,526	2011-2021 Control n = 19,183	2011-2021 Treated n = 10,786
Family history of diabetes = yes (%)		198 (2.6)	1,202 (6.3)	822 (7.6)
OGTT for fasting blood glucose = abnormal (%)		2,165 (28.8)	3,762 (19.6)	4,512 (41.8)
GDM severity = severe (%)		775 (10.3)	1,744 (9.1)	1,215 (11.3)
GDM diagnosed by the C-C criteria = yes (%)		7,323 (97.3)	17,303 (90.2)	8,419 (78.1)
Age (%)	Under 25	552 (7.3)	1,029 (5.4)	349 (3.2)
	Between 25-29	1,665 (22.1)	3,762 (19.6)	1,866 (17.3)
	Between 30-34	2,584 (34.3)	7,101 (37.0)	4,165 (38.6)
	Over 35	2,725 (36.2)	7,291 (38.0)	4,406 (40.8)
Gestational age at GDM diagnosis (mean (SD))		26.06 (6.10)	26.86 (6.02)	23.53 (7.12)
History of macrosomia = yes (%)		69 (0.9)	145 (0.8)	147 (1.4)
History of GDM = yes (%)		856 (11.4)	3,401 (17.7)	2,608 (24.2)
Parity (%)	0	3,121 (41.5)	7,611 (39.7)	3,873 (35.9)
	1	2,347 (31.2)	6,566 (34.2)	3,994 (37.0)
	more than 2	2,058 (27.3)	5,006 (26.1)	2,919 (27.1)
Pre-pregnancy BMI (%)	Underweight	100 (1.3)	391 (2.0)	76 (0.7)
	Normal	1,921 (25.5)	5,107 (26.6)	1,706 (15.8)
	Overweight	2,847 (37.8)	6,369 (33.2)	3,361 (31.2)
	Obese	2,658 (35.3)	7,316 (38.1)	5,643 (52.3)
Race/ethnicity (%)	Asian or Pacific Islander	2,919 (38.8)	8,553 (44.6)	4,560 (42.3)
	Hispanic	2,322 (30.9)	5,013 (26.1)	2,923 (27.1)
	White	1,602 (21.3)	4,090 (21.3)	2,381 (22.1)
	Black or African American	315 (4.2)	736 (3.8)	410 (3.8)
	Other or unknown	368 (4.9)	791 (4.1)	512 (4.7)
Pre-pregnancy pre-diabetes = yes (%)		479 (6.4)	1,802 (9.4)	1,915 (17.8)
Glucose challenge test value (mean (SD))		169.43 (22.38)	169.71 (22.14)	173.22 (24.32)

Balance tier	C-section delivery
1	Prognostic score quintile
2	OGTT for fasting blood glucose
	Obese pre-pregnancy BMI
3	GDM diagnosed by Carpenter-Coustan criteria
	Gestational age category at GDM diagnosis
	Pre-pregnancy pre-diabetes
	Normal pre-pregnancy BMI

Table 2.3: Balance tiers for refined covariate balance for each outcome, chosen using jointVIP plots.

Matched Design

We conduct matching with refined covariate balance using the `rcbalance` package in R and the balance tiers in Table 2.3. Post-matched jointVIP results, reflecting new levels of balance after matching, are plotted in Figure 2.2.B. For variables that were specified, post-matched biases are compared to pre-matched biases in Appendix A.6, which shows all baseline variables and summary measures to have small biases (around 0.005 or less) post-matching. Note in particular that variables with high outcome correlation are balanced especially well, a feature of the design that traditional methods based on Love plots are not equipped to guarantee.

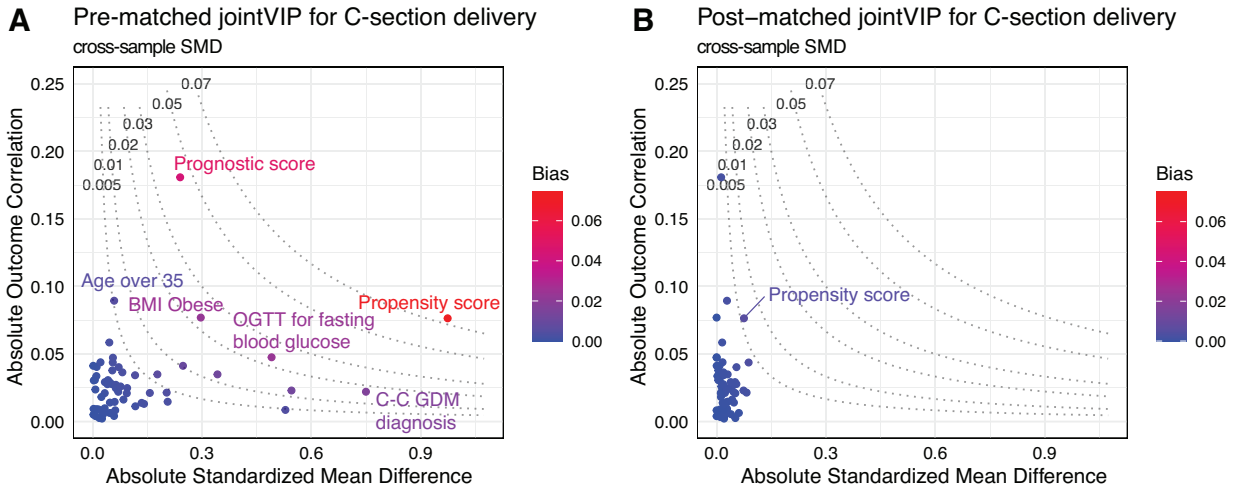


Figure 2.2: Pre-post match results for Cesarean section delivery.

2,093 treated subjects are excluded from the match due to caliper and exact matching constraints, and 8,693 pairs are matched. Those who are excluded tend to have more signs of severe GDM and higher probability of treatment; it is not surprising that it is difficult to find comparable controls for matching them (Appendix A.7). We note that the average risk difference for C-section is best understood not as an estimate of an average treatment effect on the treated[113], but as an average effect on a "marginal" population consisting of individuals for whom treatment by either arm is reasonably likely[44,68,96]. This estimand, while less common in theoretical discussions of causal inference, adheres more closely to the substantive quantity of interest for physicians who are typically more interested in guidance for patients with equipoise, and less interested in effects on patients who would clearly be assigned glyburide or not in the large majority of cases.

Outcome analysis

To perform inference, we index matched pairs by $i = 1, \cdots, I$, and individuals in each matched pair by $k = 1, 2$. For a matched pair i, one person is treated with glyburide, $Z_{ik} = 1$, and the other with MNT, $Z_{ik} = 0$, hence $Z_{i1} + Z_{i2} = 1$. Let $\mathcal{Z}$ denote the event that $Z_{i1} + Z_{i2} = 1$ for each matched pair i. Each subject ik has corresponding potential outcomes $Y_{ik}(1)$ and $Y_{ik}(0)$ for treatment with and without glyburide respectively. We collect quantities fixed in advance of treatment, including potential outcomes and covariates, in the set $\mathcal{F} = \{(Y_{ik}(1), Y_{ik}(0), \boldsymbol{x}_{ik}), i = 1, \cdots, I, k = 1, 2\}$. Our outcome of interest is a binary indicator for C-section.

We test the sharp null hypothesis, $H_0 : Y_{ik}(1) = Y_{ik}(0)$ for all i, k. Assuming that paired subjects are equally likely to receive glyburide, we can test this hypothesis by repeatedly permuting treatment indicators within pairs (independently across pairs) with probability $1/2$; this corresponds to resampling treatment indicators conditional on $\mathcal{Z}$ and $\mathcal{F}$. Since under the sharp null the outcomes remain identical regardless of treatment assignment, we can compute a test statistic under each permutation using observed outcomes and compare the actual observed value of the test statistic to this reference distribution to conduct inference. For binary outcomes, in particular, we may apply McNemar's test[78]. The above procedure relies on the assumption $Pr(Z_{ik} = 1|\mathcal{F}, \mathcal{Z}) = 1/2$ with independent assignment for each pair, which is true when unobserved confounding is absent and propensity scores are matched exactly; it is a quasi-randomization test in the sense of Zhang and Zhao[129]. In real observational studies this assumption may fail, and sensitivity analysis is needed to probe the robustness of the initial findings to such failures. We perform sensitivity analysis as described in Rosenbaum[94] Section 3.

Results

There are $2 \times 8,693$ individuals who are matched in pairs, 6,023 (34.64%) individuals delivered by C-section. Matched results are shown in Table 2.4. For control (MNT only) individuals, 33.61% delivered by C-section, and for treated (glyburide and MNT) individuals, 35.67%

		Treated with glyburide	
		C-section	not C-section
Control	C-section	1078	1844
	not C-section	2023	3748

Table 2.4: Matched analysis for Cesarean section delivery.

delivered by C-section (raw treatment-control difference of 2.06%). McNemar's test yields a one-sided p-value of 0.0020. Evaluating at significance level 0.05, there is evidence to reject the null hypothesis under a no unmeasured confounding assumption. However, the sensitivity analysis produces a threshold Γ of 1.041, which indicates that a very small degree of unmeasured confounding (the amount needed to shift a a 0.50 probability of treatment to a $1.041/(1 + 1.041) \approx 0.51$ probability of treatment) can explain away the causal effect detected. As such we find no substantial evidence that glyburide is causing the increase in cases of C-section delivery in this study.

2.4 Discussion

JointVIP is a useful tool for selecting variables to balance during the observational study design phase. One notable advantage over traditional methods is the visual ease of comparison for marginal relationships of each variable with both the outcome and treatment. Methods leveraging jointVIP can offer better bias reduction and increased robustness against unmeasured confounders[92]. Several other authors have discussed ideas closely related to jointVIP. Zhao and Yang[131] propose variable selection for fitting generalized propensity scores using measures of outcome importance and provide supporting theory suggesting the optimality of this approach. Aikens et al.[5] and Aikens and Baiocchi[4] construct an alternative design-stage visualization based partially on a pilot sample incorporating outcomes, the assignment-control (AC) plot. In contrast to jointVIP however, the AC plot represents subjects rather than variables on the plot, using the estimated prognostic score and propensity score values on the axes. AC plots and jointVIP thus provide valuable complementary representations of observational data. Finally, Cinelli and Hazlett[24] propose a similar contour plot based on omitted-variable-bias calculations that consider each variable in turn as a potential omitted confounder, for use in interpreting parameters in sensitivity analysis. For matching and weighting, the post-match jointVIP has potential to be used in a similar way. However, additional mathematical work is required to establish a mapping between the Δ_j and β_j quantities represented on the jointVIP and the parameters of existing sensitivity analysis approaches.

A natural question is why the omitted variable biases for the unadjusted bias curves should be computed under the one-covariate model in equation (2.1) instead of a model containing all measured covariates. This relates to a larger question about whether to focus on visualizing marginal measures of association between covariates and treatment or outcome, or instead to focus on conditional or partial measures that account for other variables. We focus on marginal measures rather than conditional measures (such as multiple regression coefficients from models for treatment or outcome and OVB from excluding one variable from a regression with many covariates), in contrast to previous works such as Cinelli and Hazlett[24]. While previous authors focused on post-hoc sensitivity analyses in which a single model had already been chosen for analysis, jointVIP is a pre-analysis tool aimed at helping select covariates for which to adjust. As such, it is unclear which covariates should be adjusted for in computing partial correlations with outcome and treatment. This is especially true in high-dimensional settings where the number of covariates may exceed the number of sample points in either the pilot or main analysis sample, in which case partial measures of association may not be well-defined for some sets of adjustment covariates. We also note that current standard heuristics emphasize reporting and minimizing SMDs rather than regression coefficients from a propensity score, so a marginal approach generalizes existing practice more naturally (as demonstrated above). However, developing a conditional jointVIP is an interesting topic for future work. For example, a forward-selection method with attention to multicollinearity could be developed by selecting only one variable for adjustment from the original jointVIP, then creating a conditional version of jointVIP for the remaining variables where all plotted measures adjust for the first selected variable, and iterating until a stopping criterion is reached.

While we focused on using pilot samples consisting only of controls, if extensive treatment effect heterogeneity is present this approach might underestimate the bias contributed by individual variables. Instead, one could take a pilot sample from each study arm and fit distinct treatment and control outcome correlations $\beta_j^{(1)}$ and $\beta_j^{(0)}$. A generalized version of our argument in Section 2.2 due to Zhao and Ding[130] suggests plotting $\beta_j^{(1)} p_0 + \beta_j^{(0)} p_1$ on the y-axis of the jointVIP, where p_1 and p_0 are the anticipated proportions of treated and control subjects in the final design. Of course, it may not be advisable to sacrifice treated subjects to the pilot sample for such an analysis when treatment is rare.

Another area for future work is generalizing jointVIP to allow for nonlinearity. Pearson correlation captures linear relationships but may miss strong nonlinear relationships. Nonlinear measures of importance such as the interpretable mean decrease in impurity (MDI+) derived by Agarwal et al.[2] for random forests, could in principle be used on the y-axis of the jointVIP. Two primary challenges arise. First is the question of marginal versus conditional relationships raised above, if nonlinear importance measures vary depending on the other variables included in the model. Second is the difficulty of deriving nonlinear versions of the unadjusted bias curves. Statistical interpretation of variable importance in nonlinear models such as random forest is an active research area and we are not aware of any straightforward generalization of omitted variable bias for this context.

Chapter 3

Who Is Most at Risk of Dying If Infected With SARS-CoV-2? A Mortality Risk Factor Analysis Using Machine Learning of COVID-19 Patients Over Time: A Large Population-Based Cohort Study in Mexico

In Chapter 1, we discussed variable importance in terms of causal study design. In comparison, this chapter focuses on variable importance in relation to the outcome of mortality. Two aspects are examined in detail: variable predictability and importance in an as-if causal scenario where all the individuals have the preexisting condition versus none of the individuals have the preexisting condition. For predictability, we hold other variables constant to focus on the impact of the variable of interest. For as-if scenario, we adjust for baseline characteristics to focus on the impact of the specific preexisting condition. The coauthors are Alan E. Hubbard, Juan Pablo Gutiérrez, Arturo Juárez-Flores, Kendall Kikkawa, Ronit Gupta, Yana Yarmolich, Iván de Jesús Ascencio-Montiel, and Stefano M. Bertozzi[72]. We thank the staff of C3.ai Digital Transformation Institute for their technical support and our colleagues at University of California, Berkeley, the Mexican National Autonomous University, and the Mexican Social Security Institute (IMSS) for all of the administrative and technical support that has allowed this collaboration to flourish. In addition, we acknowledge funding from C3.ai Digital Transformation Institute, National Science Foundation DGE 2146752, and Bill & Melinda Gates Foundation OPP1165144. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

3.1 Introduction

The probability of mortality associated with SARS-CoV-2 infection has varied enormously over time, among countries, and among population groups within countries[106]. Interest in understanding who is at a higher risk of death has grown as this heterogeneity became more apparent. Identifying people at higher risk of severe disease and death will help health systems better respond and focus prevention resources on protecting them. We examine Mexico, a country with a very high reported case-fatality rate (4.7%) among those who have laboratory-confirmed coronavirus disease 2019 (COVID-19) as of September 23, 2022[58]. Previous analyses in Mexico have found diabetes, obesity, hypertension, immunosuppression, and renal disease to be significant risk factors along with age and sex. Multiple authors have identified obesity and diabetes as important risk factors for mortality[11,38,81,108]. Escobedo de-la Peña et al. also found a strong association with hypertension, which is consistent with results from Giannouchos et al.[34,38]. Late-stage chronic kidney disease, although less prevalent, has also consistently been identified as a COVID-19 mortality risk factor. Older/ male patients tend to have higher mortality risks than younger/ female patients[11,34,38]. In a previous analysis, we found interactions between those comorbidities, suggesting a synergic effect when having more than one of diabetes, hypertension, and obesity (larger odds ratio when reporting the 3 conditions vs. one or two)[46]. We also found that the odds ratio increased by age group with those over age 80 having 30-fold the risk of those 20 to 29[46]. One important consideration is that the prevalence of diabetes and hypertension is positively associated with age, so it has not been clear how this interaction is related to mortality risk. A more adaptive analysis performed by Martínez-Martínez et al. developed a prediction model for severity of COVID-19, defined by hospitalization and/or mortality. They examined the relationship of 14 variables with hospitalization and mortality using interaction terms and splines to account for non-linear relationships[77]. The pattern of age, sex, and comorbidities being associated with higher mortality risk is not specific to Mexico, and the global literature on such associations is extensive. Researchers have identified old age, diabetes, obesity, chronic renal failure, and congestive heart failure to be strongly associated with severe infection amongst both sexes in the Spanish population[39]. Researchers in Brazil showed that older age, male, kidney disease, obesity and/ or diabetes are strong predictors of mortality amongst other comorbidities such as chronic liver disease, immunosuppression, and cardiovascular disease[109,126]. Another study used United Kingdom Biobank data and showed that pre-existing dementia, diabetes, chronic obstructive pulmonary disease (COPD), pneumonia, and depression were positively associated with risk of hospitalization and death[9]. An analysis from France found age, diabetes, hypertension, obesity, cancer, and kidney and lung transplants to be associated with risk of COVID-19-related hospitalization and mortality, among others[105]. A Canadian study reported dementia, chronic kidney disease, cardiovascular disease, diabetes, COPD, severe mental illness, organ transplant, hypertension, and cancer to be significant predictors of mortality[37]. Studies presented here is a non-exhaustive list of research studying COVID-19 risk factors and mortality. Recent meta-analyses and systematic reviews find significant mortality attributed to these pre-existing conditions[6,64,101,106]. Our goal in this

study is not only to predict mortality using demographic factors and comorbidities, but to show how those predictions change over time in this rapidly evolving pandemic. Although mortality risk estimation and risk factor identification have been examined in prior studies, we are concerned about the statistical validity and interpretation of the standard methods. A commonly used prediction tool, logistic regression, assumes a linear relationship of predictors against the log odds of mortality risk, but this logit-linear assumption will lead inevitably to biased estimates of risk (either under- or over-predict the risk) for subsets of the population. We instead used flexible, data-adaptive methods that can capture non-linearities in the dose-response, such as potential nonlinear interactions between the predictors (e.g., the potential interaction of age and diabetes on predicting death)[88,120]. The better the model fits the study population; the more likely estimates are closer to the true joint relationship of mortality and risk factors. We included pre-existing conditions, demographic variables, the Mexican state where the patient was treated, and the month that the patient initiated care to fit our prediction algorithm. We conducted the analysis using an ensemble machine learning algorithm, super learner, to form optimal combination of predictions from multiple machine learning methods[88,120]. We also estimated the comparative importance of variables for mortality risk prediction (holding all other variables constant) by nonparametrically estimating quantities inspired by causal parameters (parameters that compare so-called counterfactual distributions, in our case, causal relative risks). The statistical goal is to estimate and provide robust inference for impact estimates of the predictors without the arbitrary modeling assumptions that characterize the great majority of prior work[84].

3.2 Methods

Study population and design

The study population is drawn from the Mexican Social Security Institute (IMSS), a vertically integrated insurance and health system that provides coverage for over 60 million private sector employees and their families, including their parents, children and spouse. IMSS also provided care as part of the COVID-19 response for some non-beneficiaries, who are also included in the dataset. The data were recorded from March 1st, 2020, to November 3rd, 2021 in a platform called SINOLAVE. They reflect the entire population of 4,482,292 patients who were registered as receiving care for suspected COVID-19 at an IMSS facility. The dataset and the data entry process have been described previously[63]. The demographic variables include age, sex, insured by IMSS, and indigenous status. The data contains pre-existing conditions reported by the patient or the family at presentation: asthma, cardiovascular disease, chronic liver disease, chronic obstructive pulmonary disease, diabetes, hemolytic anemia, human immunodeficiency virus, hypertension, immunosuppression, neurological disease, obesity, cancer, renal disease and tuberculosis, as well as whether the patient currently smokes. Patients were asked at presentation about their pre-existing health conditions; these were not ascertained with reference to the patient's medical record, even for those

	All time (2020/03-2021/11)	Phase 1 (2020/03-2020/10)	Phase 2 (2020/11-2021/03)	Phase 3 (2021/04-2021/11)
Sample size	1,423,720	303,278	425,698	694,744
Demographic variables				
Age in years (mean (SD))	42.15 (15.70)	46.41 (16.04)	44.89 (16.27)	38.61 (14.34)
Sex = male (%)	729,782 (51.3)	158,248 (52.2)	218,165 (51.2)	353,369 (50.9)
Insured by IMSS = yes (%)	1,358,440 (95.4)	288,588 (95.2)	402,754 (94.6)	667,098 (96.0)
Indigenous = yes (%)	7,381 (0.5)	2,200 (0.7)	1,628 (0.4)	3,553 (0.5)
Pre-existing conditions				
Hypertension = yes (%)	228,901 (16.1)	72,615 (23.9)	83,735 (19.7)	72,551 (10.4)
Diabetes = yes (%)	169,869 (11.9)	55,551 (18.3)	61,120 (14.4)	53,198 (7.7)
Obesity = yes (%)	181,736 (12.8)	55,965 (18.5)	60,217 (14.1)	65,554 (9.4)
Smoking = yes (%)	87,161 (6.1)	21,253 (7.0)	28,346 (6.7)	37,562 (5.4)
Asthma = yes (%)	25,297 (1.8)	7,951 (2.6)	7,765 (1.8)	9,581 (1.4)
Renal Disease Diagnosis = yes (%)	24,099 (1.7)	8,912 (2.9)	8,555 (2.0)	6,632 (1.0)
Outcome				
Death = yes (%)	149,805 (10.5)	53,530 (17.7)	62,517 (14.7)	33,758 (4.9)

Table 3.1: Summary table of baseline variables and pre-existing conditions

patients insured by the IMSS. The data also includes the Mexican state in which the patient received care, COVID-19 test results (from both polymerase chain reaction (PCR) tests and antigen tests), the month that the patient initiated care, and COVID-related mortality. The outcome, death, is ascertained as COVID-related mortality within this study period between March 2020 and November 2021; we only consider deaths after patients initiated care. In addition, we extracted a different dataset from the National Council of Science and Technology to determine the dominant circulating variant in each month[29]. A short summary can be found in Table 3.1 (Appendix B.1). We define COVID-19 positive as a positive PCR or antigen test.

From the full data set, we generated an analytic sample (n = 1,423,720) (Appendix B.2). We exclude those under the age of 20 years, those without any positive COVID-19 test result from either the PCR or antigen tests, and those with unknown pre-existing conditions. We also create a phase variable that corresponds to changes in the epidemic curve into three: phase 1 is from March 1st, 2020, to October 31st, 2020, phase 2 is from November 1st, 2020, to March 31st, 2021, and phase 3 is from April 1st, 2021, to November 3rd, 2021 as previously described[63].

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Statistical analysis

Mortality risk prediction using super learner (SL)

We predict mortality risks with SL[88,120], using predictors: pre-existing conditions, demographic variables, the Mexican state where the patient was treated, and the month that the patient initiated care. SL combines a set of user-supplied machine learning algorithms, which includes both simple, parametric fits and flexible algorithms, to create an optimally-weighted combination. This optimal fit is found by creating a combination of algorithms that minimize the cross-validated risk (in our case, the negative log-likelihood). SL has the property that asymptotically it will perform at least as well as the best fitting algorithm in the library[88,120]. Thus, it is important to include a diverse and large set of learners as candidates to ensure the model can fit complex patterns if warranted, but also, simpler, parametric models if simpler fits are sufficient. The following learners were included in the SL library: Bayesian additive regression trees[23], Bayesian generalized linear model[31], elastic net regression[47], empirical mean, generalized additive model[74], least absolute shrinkage and selection operator regression[116], logistic regression, multivariate adaptive regression splines[36], random forest[17], ridge regression[56], and extreme gradient boosting algorithms[21]. We estimate the prediction performance, via the AUC, and derive a 95% confidence interval for the estimated AUC[66]. We compare the SL fit using all predictors listed above to a logistic regression with only age entered as a linear term. We compute the AUC for the resulting SL/logistic regression fits with 3-fold cross validation on the 80%, both on the same data used to estimate SL/logistic regression models (training AUC), as well as a more realistic assessment by using the test set – the left-out 20% of the available data (testing AUC). To interpret the final prediction model generated by the SL fit, we use the permutation-based variable importance measure to identify variables that influence the SL model's prediction[17]. This is performed by permuting the predictor variables one at a time (keeping the other variables fixed) and measuring the magnitude of the decline on the predictive performance (as measured by the change in the average negative log-likelihood). This provides a list of variables ranked by the relative importance to prediction fit but does not provide information on the variable impact on mortality, which led us to another measure of relative risk (RR) using targeted maximum likelihood estimation (TMLE).

Pre-existing condition relative risk estimate through targeted maximum likelihood estimation

For pre-existing conditions, we estimated a different variable importance measure that is not focused on prediction accuracy but on estimating potential impacts of pre-existing conditions on mortality risk. The impact is estimated by the RR of adjusted means (adjusted for baseline confounders) for the population if everyone had the specific pre-existing condition of interest (the numerator) versus the same population where no one has the specific pre-existing condition (the denominator). To estimate RRs, we used cross-validated targeted minimum-loss-based estimation (cross-validated TMLE). TMLE is a semiparametric,

	All time (2020/03-2021/11) AUC (95% CI)	Phase 1 (2020/03-2020/10) AUC (95% CI)	Phase 2 (2020/11-2021/03) AUC (95% CI)	Phase 3 (2021/04-2021/11) AUC (95% CI)
Super learner fit	Training: 0.916 (0.915-0.917) Testing: 0.907 (0.905-0.908)	Training: 0.887 (0.885-0.888) Testing: 0.873 (0.870-0.876)	Training: 0.904 (0.903-0.906) Testing: 0.895 (0.892-0.897)	Training: 0.914 (0.913-0.916) Testing: 0.906 (0.902-0.909)
Age only logistic regression fit	Training: 0.874 (0.873-0.875) Testing: 0.874 (0.872-0.876)	Training: 0.845 (0.843-0.846) Testing: 0.846 (0.842-0.850)	Training: 0.868 (0.866-0.870) Testing: 0.871 (0.868-0.874)	Training: 0.867 (0.865-0.869) Testing: 0.871 (0.866-0.875)

Table 3.2: Prediction results.

substitution estimator that has shown to be asymptotically efficient (unlike the inverse probability of treatment-weighting estimators[90]). It also has some robustness advantages over other semiparametric efficient approaches, such as augmented inverse probability weighting. TMLE estimates parameters that, under certain assumptions, can be interpreted as potential causal impacts of these factors on mortality, in our case, in the form of a causal relative risk. Our ensemble machine learning is optimized for prediction, but it does not directly provide measures of individual variable importance. We augmented our prior SL analysis using the TMLE to generate interpretable estimates of variable impact with robust standard errors[89,118,119]. Both analyses using SL and TMLE are conducted in programming language R; the code used to conduct this analysis is publicly available on GitHub (link: https://github.com/ldliao/mexPred).

3.3 Results

Descriptive results show the age distribution of laboratory-confirmed patients across the three different epidemic phases (Appendix B.3). Phases 1 and 2 have similar distributions, and there are more young people (under 30) in phase 3. The six most prevalent pre-existing conditions are hypertension, obesity, diabetes, smoking, asthma, and renal disease (Appendix B.4). The prevalence of all pre-existing conditions decreased over the three phases, and prevalence of hypertension, obesity, and diabetes were drastically reduced in phase 3.

Super learner (SL) prediction

SL fit has high prediction accuracy on the testing set (AUC: 0.907 (95% CI: (0.905-0.908)) (Table 3.2). The SL fit leverages multiple machine learning models: the XGBoost models, generalized additive model, and random forest for prediction (Appendix B.5). The simple logistic regression has a lower AUC (testing AUC: 0.874 (95% CI: (0.872-0.876)) than the

SL fit, as expected, as it only uses age as a predictor (Table 3.2). However, the simple model is already highly predictive, and the difference is small yet significant. The logistic regression model overpredicts mortality risks for those roughly above age 75 compared to the SL prediction (Figure 3.1). In line with the simple age-only logistic regression model, permuted variable importance on the SL fit shows, while holding other variables constant, age is consistently the most important for SL prediction in average mortality risk (Appendix B.6 and B.7). Having multiple comorbidities can dramatically increase risk for those individuals (Figure 3.2).

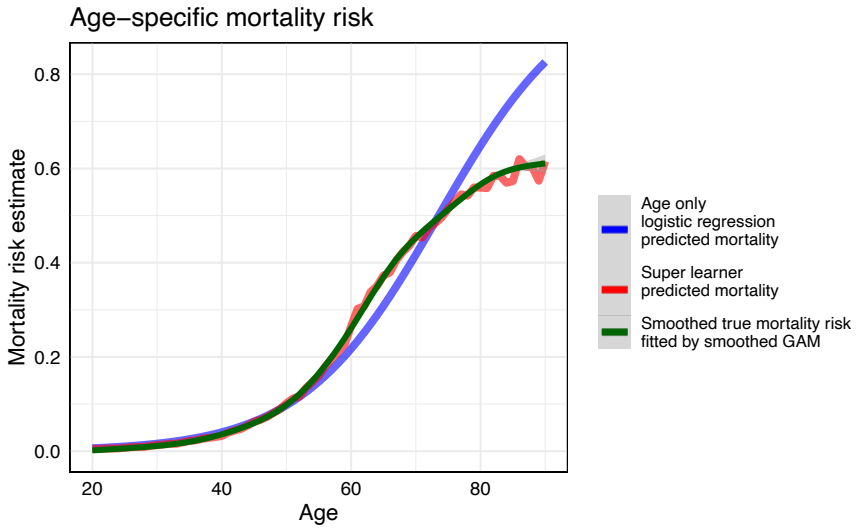


Figure 3.1: Mortality risk prediction comparing age only logistic regression and super learner.

Relative risks of pre-existing conditions

To assess the impact of each pre-existing condition, we estimate their respective relative risks (RRs) of mortality, adjusting for demographic variables. We report the estimated RRs in Table 3, ordered by impact (most to least) (Appendix B.8). The RRs compare the expected risk if all patients have the pre-existing condition (with) versus if all patients do not have the condition (without). The highest impact pre-existing condition is renal disease (RR: 3.783, 95% CI: (3.705, 3.862)); diabetes, obesity, and hypertension also have high impact individually (RR: 1.432-1.847). Minimal differences between the risk estimates are shown for smoking and asthma (RR: 1.049 and 1.037, respectively).

%% table

The phase analyses indicate pre-existing conditions are especially important in phase 3. Phase 1 and 2 are very similar in terms of both risk prediction and adjusted mortality risk estimates. However, in phase 3, age is less important in prediction (Appendix B.7) and RRs

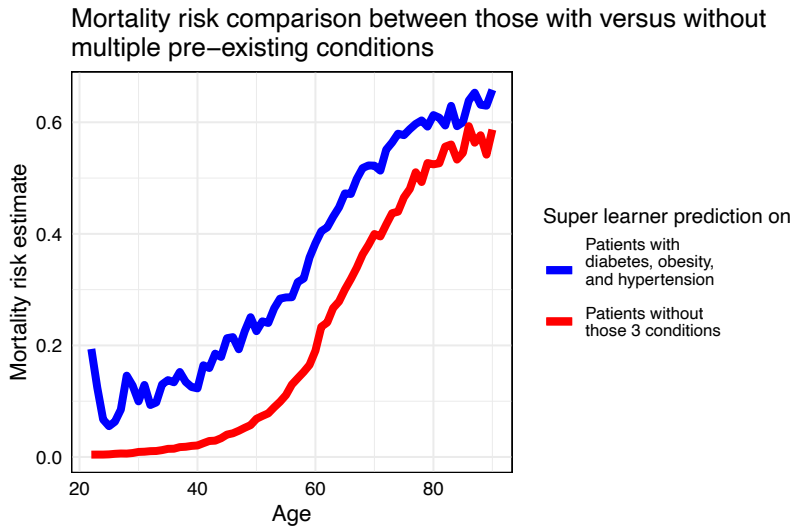


Figure 3.2: Super learner predicted mortality risk averaged by specific age in two subgroups: those having all obesity, diabetes, and hypertension pre-existing conditions versus those without.

drastically increase for every comorbidity. The adjusted risks show the decrease for each pre-existing condition in phase 3 (Appendix B.9).

3.4 Discussion

Our analysis of (> 1.4 million) laboratory-confirmed COVID-19 patients demonstrates that age is by far the most important predictor of average mortality. For those patients with renal disease, diabetes, hypertension, or obesity, having the comorbidity further increases their risk of mortality. A patient with diabetes, hypertension, and obesity is roughly comparable to a patient 20 years older with none of the conditions, based on the predicted mortality (Figure 3.2). Thus, having a comorbidity increases risk of mortality and should be considered at any age. The reason that comorbidities add little to the predictive power at younger ages is that hypertension and diabetes are age-related and the reported onset is often for those over 30, so the pre-existing conditions are far less prevalent. Our prediction results using machine learning methods predict better than previous studies, and we demonstrated the feasibility and robustness of using machine learning methods targeted for prediction and variable impact. SL model prediction has an AUC of 0.907, which is higher than any previous Mexican study (AUCs from 0.634 to 0.824)[77,125]. Although age has been well reported by previous studies as important [6, 38, 39], our analysis is more robust because we do not assume a pre-specified functional relationship between the explanatory variables and the predicted variable, and thereby avoid any arbitrary groupings into age categories. Moreover,

since those above age 60 have a higher prevalence of comorbidities, relying on simple logistic regression models can greatly overpredict the average mortality risk for the elder patients. Our study applies TMLE to estimate the adjusted mortality risk ratios for each comorbidity to provide more robust impact estimates that respect time ordering and account for background variables. We find consistent results of comorbidities compared to previous studies, and present phase analyses highlighting the changes in relative risks over time. Previous results from logistic regressions indicated odds ratios of 1.458-2.48 for renal disease, 1.237-1.74 for diabetes, 1.173-1.47 for obesity, 1.194-1.315 for hypertension, 0.852-1.02 for smoking, and 0.74-1.420 for asthma[54,82,125]. Although our analysis is generally consistent with previous findings, our RR estimations have less uncertainty. Renal disease has the greatest impact on mortality, followed by diabetes, hypertension, and obesity; smoking and asthma have negligible impact on mortality risk. This phase-specific analysis produced a seemingly paradoxical finding. The impact of comorbidities on predicted mortality decreased with time (primarily between the second and third wave), but the RR on mortality dramatically increased for the same conditions (Appendix B.8 and B.9). The apparent explanation is that mortality risk for people without the comorbidities fell faster than for people with them, increasing the relative risk. The decrease in mortality risk is multifactorial and includes a decrease in susceptibility over time (due to prior infection and vaccination), improved treatment, enhanced healthcare response and opportunity to be admitted to a hospital or ICU, and less virulent viral subtypes. This implies that as herd immunity increases, medical resources should focus even more on protecting vulnerable people at older age and those with comorbidities since they are even more likely to experience severe outcomes compared to those who are younger and/or healthier. Readers should be cautious about extrapolating our findings to other populations. Although our sample is large and includes patients from all parts of Mexico, most of the patients were IMSS beneficiaries. In order to access IMSS health services, patients require: a) be a formal-sector worker or retired, b) be a direct dependent of such an employee, c) be a bachelor or postgraduate student in a public institution, d) voluntarily enroll by paying a fee. Thus, the IMSS population skews toward the upper half of the income distribution. Populations without similar access to health services may have different results. It is also important to consider the potential impact of data quality. Pre-existing conditions were self-reported and likely also inconsistently recorded, perhaps in systematic ways that could have biased the results. For example, if people with severe diabetes were more likely to report diabetes as a pre-existing condition, we may overestimate the impact of diabetes on mortality. It is also important to consider what predictive variables are included in this model. We sought to predict risk for an individual in the population using their characteristics prior to infection. In other words, what is this person's risk of death from COVID-19 if they were to be infected? The answer to this question best informs the question of who should be prioritized for protection against infection or for early therapeutic interventions following infection. It does not attempt to predict the likely mortality of a patient who presents to the health services with COVID-19 because information about that patient's severity of their COVID-19-related symptoms will represent important additional predictors of their mortality risk.

Chapter 4

Prognostic Adjustment With Efficient Estimators to Unbiasedly Leverage Historical Data in Randomized Trials

In Chapter 2, we utilized the previously collected (external, pilot) data to inform study design. In Chapter 3, we investigated the predictive performance and as-if causal importance of mortality risk factors using machine learning. In this chapter, previously collected external data, referred to as historical data, is used in conjunction with machine learning to improve trial analysis in the form of prognostic covariate adjustment. The coauthors are Emilie Højbjerre-Frandsen, Alan E. Hubbard, and Alejandro Schuler. We would like to thank study participants and staff for their contributions. This research was conducted on the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley. This computing resource was supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer. The authors thank Christopher Paciorek for answering Savio related inquiries. This research was made possible by funding from the National Science Foundation DGE 2146752 and global development grant OPP1165144 from the Bill & Melinda Gates Foundation. This research also received funding from Innovation Fund Denmark (Grant number 2052-00044B) to Novo Nordisk A/S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.1 Introduction

Practical, financial, and ethical concerns often preclude large randomized trials, which limits their power[15,41,115]. On the other hand, historical (often observational) data are often plentiful, and there are many existing methods for including historical data in trial analyses in order to boost power[127]. "Data fusion" methods simply pool trials with historical data[10,26,107]. Bayesian methods, which naturally rely on assumptions in the form of specified

priors from historical data, are also popular in the literature[55,69]. Similar problems have also been addressed in the generalizability and transportability research[30,61,107]. Recent studies proposed machine learning methods to integrate prior observational studies into trial analyses[67,70]. While pooled estimators, which integrate historical data with trial data, are an active area of research (for example, Dang et al.[28]), our focus is solely on improving trial analysis.

Unfortunately, the aforementioned approaches that rely on validity of historical data are all sensitive to unobservable selection biases and must therefore be used with extreme care. The fundamental problem is that the historical population may differ systematically from the trial population in ways that impact both treatment assignment and outcome. For example, if the historical population did not have access to a modern standard-of-care, adding historical controls would artificially make any new drug seem more effective than it really is. Observable differences in populations can potentially be corrected under reasonable assumptions, but shifts in unobserved variables are impossible to detect or correct.

We take the approach of covariate adjustment to increase efficiency. In recognition of using covariates to reduce estimation uncertainty, the U.S. Food and Drug Adminstration recently released guidance on adjusting for covariates in randomized clinical trials[35]. See Van et al. (2023) for summarized methods using covariate adjustment[122]. Our research builds on Schuler et al. (2021), who suggest using the historical data to train a *prognostic model* that predicts the outcome from baseline covariates[104]. They then adjust for the model's predictions on the trial data in the trial analysis using linear regression, namely the "prognostic adjustment". A similar research proposed by Holzhauer and Adewuyi (2023) recommends using a "super-covariate," combining multiple prognostic models into a single covariate for adjustment[57]. However, both these methods are limited to trial analyses using linear regression models.

Our task in this paper is to extend the prognostic adjustment approach beyond linear regression, specifically, to "semiparametrically efficient" estimators. Semiparametrically efficient estimators are those that attain the semiparametric efficient variance bound, which is the smallest asymptotic variance that any estimator can attain. The use of efficient estimators thus tends to reduce the uncertainty of the treatment effect estimate. These estimators leverage machine learning internally to estimate the treatment or the outcome model, or both; for example, the augmented inverse probability weighting estimator (AIPW) and the targeted maximum likelihood estimator (TMLE) are commonly used to evaluate the average treatment effect[22,32,42,121?] . These estimators have been shown to improve the power of trials over unadjusted or linearly adjusted estimates[98].

In this study, we aim to improve power even further by incorporating historical data via prognostic adjustment. Our approach guarantees asymptotic efficiency of the trial treatment effect and more importantly, promises benefits in finite-sample efficiency and robust inference.

4.2 Framework and notation

We follow the causal inference framework and roadmap from Petersen and van der Laan[84]. First, we define each observational unit $i \in \{1, \ldots, n\}$, as an independent, identically distributed random variable, O_i with true distribution P. In our setting, each random variable $O = (W, A, Y, D)$ contains associated p baseline covariates $W \in \mathbb{R}^p$, a binary treatment indicator A, denoting whether a unit is in the control group ($A = 0$), or in the treatment group ($A = 1$), an outcome Y, and an indicator D denoting whether a unit is in either the trial ($D = 1$) or historical ($D = 0$) data sample.

We will assume that the trial data is generated under the setting of an RCT, such that $P(A = a|W, D = 1) = \pi_a$, with some positive constant π_a denoting the treatment probability for $a \in \{0, 1\}$. Define $\mu_a(W) = E_P[Y|A = a, W, D = 1]$ as the conditional outcome means per treatment arm in the trial. Let $\rho_d(W) = E[Y|W, D = d]$ denote the *prognostic score* for a dataset d[49]. When referenced without subscript (ρ) we are referring to the prognostic score in the historical data $D = 0$.

The fundamental problem of causal inference comes from not being able to observe the outcome under both treatment types. We know that for each individual, $Y = Y^A$, i.e., we observe the potential outcome corresponding to the observed treatment. To calculate the causal parameter of interest, we define the (unobservable) causal data to be (Y^1, Y^0, A, W, D), generated from a causal data generating distribution P^*. In this study, we are interested in the causal average treatment effect (ATE) in the trial population:

$$\Psi^* = E_{P^*}[Y^1 - Y^0|D = 1]$$

which due to randomization in the trial is equal to the observable quantity:

$$\Psi = E_P[\mu_1(W) - \mu_0(W)|D = 1]$$

where $\mu_a(W) = E_P[Y|A = a, W, D = 1]$ is the conditional mean outcome in treatment arm $a \in \{0, 1\}$ from the observable data distribution.

Let $(\mathbf{W}, \mathbf{Y})$ denote a dataset with observed outcome $\mathbf{Y} = [Y_1, \ldots, Y_n]$ and the observed covariates $\mathbf{W} = [W_1, \ldots, W_n]$, where $(Y, W) \in (\mathbb{R} \times \mathbb{R}^m)$. Furthermore, let $\mathcal{L} : (\mathbf{W}, \mathbf{Y}) \mapsto f$ denote a machine learning algorithm that maps $(\mathbf{W}, \mathbf{Y})$ to a learned function f that estimates the conditional mean $E[Y|W]$. The algorithm $\mathcal{L}$ may include detailed internal model selection and parameter tuning, and the algorithm works with predictors and data of any dimension (i.e., m, n are arbitrary). Let $\tilde{\mathbf{Y}}, \tilde{\mathbf{W}}$ represent the historical dataset of size $\tilde{n}$, which is a draw from $P^{\tilde{n}}(Y, W|D = 0)$. We use $\hat{\rho}_0 = \mathcal{L}(\tilde{\mathbf{Y}}, \tilde{\mathbf{W}})$ (or just $\hat{\rho}$) to refer to an estimate of prognostic score learned from the historical data. Let $(\mathbf{Y}, \mathbf{A}, \mathbf{W})$ represent the trial data set of size n, which is a draw from $P_n(Y, A, W|D = 1)$. In a slight abuse of notation, let $\hat{\psi} = \hat{\psi}(\mathbf{Y}, \mathbf{A}, \mathbf{W})$ denote the mapping between trial data and our estimate $\hat{\psi}$ using an efficient estimator. For example, $\hat{\psi}$ could denote the cross-fit AIPW estimator described in Schuler (2021)[102].[?]

4.3 Method

Efficient estimators with prognostic score adjustment

Our proposed method for incorporating historical data with efficient estimators is simple: we first obtain a prognostic model by performing an outcome prediction to fit the historical data ($D = 0$) using a machine learning algorithm $\hat{\rho} = \mathcal{L}(\tilde{\mathbf{Y}}, \tilde{\mathbf{W}})$. We then calculate the value of the prognostic score in the trial by feeding in all units' baseline covariates: $\mathbf{R} = \hat{\rho}(\mathbf{W})$. This prognostic score can be interpreted as a "pre-learned" dimension reduction for the trial covariates. Lastly, we estimate the ATE from the trial data, augmented with the prognostic score as an additional covariate, using an efficient estimator: $\hat{\psi}(\mathbf{Y}, \mathbf{A}, [\mathbf{W}, \mathbf{Y}])$.

In practice, we suggest using a cross-validated ensemble algorithm (also called "super-learner") for $\mathcal{L}$[120]. The super learner is known to perform as well as the best machine learning algorithm included in the library[88,120]. The library in the super learner should include a variety of nonparametric and parametric learners, such as gradient boosting, random forest, elastic net, and linear models[88,120].

For an efficient estimator, adding a fixed function of the covariates as an additional covariate will not change the asymptotic behavior[80,99]. Thus our approach will never be *worse* than ignoring the historical data (as it might be if we pooled the data to learn the outcome regression). However, it also means that our approach cannot reduce asymptotic variance (indeed it is impossible to do so without making assumptions).

Nonetheless, we find that the *finite-sample* variance of efficient estimators is far enough from the efficiency bound that using the prognostic score as a covariate generally decreases the variance (without introducing bias) and improves estimation of the standard error. Mechanistically, this happens because the prognostic score "jump-starts" the learning curve of the outcome regression models such that more accurate predictions can be made with fewer trial data. This is especially true when the outcome-covariate relationship is complex and difficult to learn from a small trial. A "small" trial refers to a smaller sample size than traditionally needed for an unadjusted estimator when calculating the desired power. It is well-known that the performance of efficient estimators in RCTs is dependant on the predictive power of the outcome regression. Therefore improving this regression (by leveraging historical data) can reduce variance.

We expect finite-sample benefits as long the trial and historical populations and treatments are similar enough. But even if they are not identical, the prognostic score is still likely to contain very useful information about the conditional outcome mean.

In the following subsections 4.3 to 4.3, we theoretically show how adjusting for a prognostic score with an efficient estimator can improve estimation in a randomized trial. The implications are that small-sample point estimation and inference should be improved even though efficiency gains will diminish asymptotically. In an asymptotic analysis where the historical data grows much faster than the trial data, we show that using the prognostic score speeds the decay of the empirical process term in the stochastic decomposition of our estimator. The implications are that finite-sample point estimation and inference should be

improved even though efficiency gains will diminish asymptotically. The material assumes expertise with semiparametric efficiency theory and targeted/double machine learning. We present our results at a high level and do not present enough background for a casual reader. Two good starting points for this material are Schuler and van der Laan and Kennedy[65,103]. The casual reader may skip this section and if they are comfortable with the above heuristic explanation of why prognostic adjustment may improve performance.

No asymptotic efficiency gain

Before showing how adjusting for a prognostic score for an efficient estimator can benefit estimation, we show that adding any prognostic score to an efficient estimator *cannot* improve asymptotic efficiency. To see this, we will start by considering the counterfactual means $\psi_a = E[E[Y|A = a, W]] = E[\mu_a(W)]$ for any choice of $a \in \{0, 1\}$. We will return to the ATE shortly, but for now, it will make our argument clearer to only consider counterfactual means. Consider any efficient estimator for ψ_a in a semiparametric model (known treatment mechanism) over the trial data (Y, A, W). The *influence function* of an estimator completely determines its asymptotic behavior. By definition, any efficient estimator of $E[\mu_a(W)]$ must have an influence function equal to the canonical gradient, which is referred to as the *efficient* influence function:

$$\phi_a(Y, A, W) = ((I(A = 0))/\pi_a)(Y - \mu_a(W)) + (\mu_a(W) - \psi_a)$$

where I is the indicator function and $\pi_a = P(A = a)$ is the fixed, known propensity score. Consider now a distribution over $(Y, A, [W, R])$, where $R = g(W)$ for any fixed function g playing the role of a prognostic model. The efficient influence function in this setting is the same as the above except $\mu_a(W) = E[Y|A = 0, W]$ is replaced by $\mu_a(W, R) = E[Y|A = 0, W, R]$. But since the prognostic score $R = g(W)$ is a fixed function of the covariates W, conditioning on the prognostic score after conditioning on W does nothing, and we obtain $\mu_a(W, R) = \mu_a(W)$. Therefore, the prognostic score does not change the influence function and therefore cannot improve asymptotic efficiency. This holds even if you consider a random prognostic score learned from external data. The same holds for the ATE parameter since the efficient influence function in this case is $\phi_{ATE} = \phi_1 - \phi_0$.

The fundamental issue is that the asymptotic efficiency bound cannot be improved without considering a different statistical model, e.g. distributions over (Y, A, W, D). The problem is that in considering a different model we must also introduce additional assumptions to maintain identifiability of our trial-population causal parameter. For example, Li et al. consider precisely this setup and rely on an assumption of "conditional mean equivalence" $E[Y|A, W, D = 1] = E[Y|A, W, D = 0]$ to maintain identification while improving efficiency[70]. Similarly, an analysis following Chakrabortty et al. shows that efficiency gains are also possible if we assume the covariate distributions are the same when conditioning on D[20]. In this paper, we take the covariate adjustment approach, incorporating the external data without these explicit assumptions and therefore we have to look for benefits in finite sample improvements.

Improving point estimation

To understand the benefits of prognostic adjustment we must consider the non-asymptotic behavior of our estimator. As above we will consider estimation of the treatment-specific mean ψ_a, but from now on we will omit the a subscripts to reduce visual clutter.

Consider the following decomposition:

$$\hat{\psi} - \psi = P_n\phi + (P_n - P)(\hat{\phi} - \phi) + S - P_n\hat{\phi}$$

This sort of decomposition is common in the analysis of efficient estimators[65,103]. As above, ϕ here denotes the efficient influence function of ψ. We use the empirical process notation $P_n\hat{\phi} = n^{-1}\sum_{i=1}^{n}\hat{\phi}(Y_i, A_i, W_i)$ to denote a sample mean, $P\hat{\phi} = E[\hat{\phi}(Y, A, W)]$ to denote a population mean (not averaging over randomness in $\hat{\phi}$), and $\hat{\phi}$ a plug-in estimate of ϕ with μ estimated by regression. We will analyze each term and see what difference adjusting for the prognostic score makes in this general decomposition. Note that for an efficient estimator, the last term $-P_n\hat{\phi} = 0$ by construction. Eliminating this "plug-in bias" term is the purpose of bias correcting schemes such as TMLE or efficient estimating equations[65,103].

Of the remaining terms, the first is the efficient influence function term, $P_n\phi$. We have already shown that an efficient estimator leveraging the prognostic score has the same influence function as one without. It is also known that the remainder term $S = [\hat{\phi} - \phi] + P\hat{\phi}$ can be bounded by the product of estimation errors in the outcome and propensity regressions $||\hat{\mu} - \mu||||\hat{\pi} - \pi||$[65,103]. This is exactly zero in a randomized trial since the propensity score is known. Therefore both of these terms are unaffected by prognostic adjustment.

That leaves us with only the "empirical process" term $(P_n - P)(\hat{\phi} - \phi)$. With cross-fitting this term can be shown to be $O_P(||\hat{\phi} - \phi||n^{-1/2})$[65,103]. In our setting, the estimated influence function depends only on the estimated outcome regression and we thus have $(P_n - P)(\hat{\phi} - \phi) = O_P(||\hat{\mu} - \mu||n^{-1/2})$. This is $o_P(n^{-1/2})$ in all cases where the regression is L_2 consistent, but the actual rate *can be faster* depending on how quickly $||\hat{\mu} - \mu||$ converges. We'll use t_n to denote this rate, i.e., $||\hat{\mu} - \mu|| = o_P(t_n)$. This is the mechanism by which prognostic adjustment improves efficient estimators. When the prognostic score is used to estimate $\hat{\mu}$, the norm $||\hat{\mu} - \mu||$ is generally smaller than it otherwise would be.

We can formalize this asymptotically. Recall that n denotes the trial sample size and $\tilde{n}$ denotes the historical sample size. Consider a historical sample much larger than the trial in the limit: $n/\tilde{n} = r_n \to 0$. For example, presume $\tilde{n} = n^2$ in which case $r_n = 1/n$. Presume that the historical distribution of covariates and outcome is the same as that in the trial control arm, or simply that $\rho(w) = \mu_0(w)$ (this is a best-case scenario). Presume we have a learning algorithm $\mathcal{L}$ that in some large nonparametric function class can learn functions in an L_2 sense at rate t_n. Let $\hat{\rho}_{\tilde{n}}$ be the prognostic score learned from the $\tilde{n}$ historical samples using this learner.

Instead of fitting our trial outcome regression with the prognostic score as a covariate, presume that we directly take $\hat{\mu}_{(n,0)} = \hat{\rho}_{\tilde{n}}$. In other words, we use our prognostic model as the control outcome regression in the trial, ignoring the trial data. That's sensible in this hypothetical setting because 1) the prognostic model will indeed converge to the true

outcome function and 2) the amount of omitted trial data is vanishingly small relative to the historical data. We can also interpret this as an estimator of $E[Y|A = 0, W, \rho_{\tilde{n}}(W)]$ in the trial data (i.e., our prognostically adjusted outcome regression): we should expect that when the prognostic model is good, our learner will simply return the prognostic score untouched to model the control counterfactual outcome. Or, more formally, if a parametric learner is used on top of the prognostic score (e.g. a working model $Y = \beta_0 + \hat{\rho}(W)$) then the total error will be composed of a root-n term from learning the parameters and a higher order empirical process term. If we used our learner to fit the outcome regression from trial data alone, our rate on the L_2 norm of the outcome regression would be t_n, e.g. $n^{-1/10}$ for example. However, if we use the prognostic score the rate is $t_{\tilde{n}}$ ($\tilde{n}$ instead of n), e.g. $\tilde{n}^{-1/10}$. But recalling $n/\tilde{n} = o(r_n)$ ($r_n = 1/n$, for example), we can express $t_{\tilde{n}}$ as $(t_n r_n)/n$. Essentially we can plug in $\tilde{n} = n^2$ into $t_{\tilde{n}} = \tilde{n}^{-1/10}$ to see $t_n = (n^2)^{-1/10} = n^{-1/5}$. This can dramatically increase the speed at which $||\hat{\mu} - \mu|| \to 0$ in terms of n and consequently affect the rate of convergence of the empirical process term, making it decay faster than it would without the prognostic adjustment.

The result of making the empirical process term higher order is to reduce finite-sample variance of our point estimate. With cross-fitting the empirical process term is exactly mean-zero[65,103], so finite-sample bias is unaffected.

Improving standard error estimation

We can apply similar arguments to show that performance of the plug-in estimate of asymptotic variance $P_n\hat{\phi}^2$ (based on our estimated influence function $\hat{\phi}$) is also improved by prognostic adjustment. The difference between the estimate and the true asymptotic variance $P\phi^2$ can be decomposed as

$$P_n\hat{\phi}^2 - P\phi^2 = (P_n - P)\phi^2 + (P_n - P)(\hat{\phi}^2 - \phi^2) + P(\hat{\phi}^2 - \phi^2)$$

The first term here is a nice empirical mean which by the central limit goes to zero at a root-n rate and which is unaffected by prognostic adjustment. The second term is similar to the empirical process term discussed above in the context of point estimation and by identical arguments this term decays faster when prognostic adjustment is used (note L_2 convergence of $\hat{\phi}$ implies the same for $\hat{\phi}^2$ under regularity conditions that are satisfied in our setting because π is a known constant bounded away from 0). This term is always higher-order than $n^(-1/2)$ and thus asymptotically negligible, but possibly impactful in small, finite samples. It is also mean-zero. Together, this means that its improved rate with prognostic adjustment translates to less finite-sample variability in our estimate of the standard error.

The last term is bounded by $||\hat{\phi}^2 - \phi^2||$ so this term also decays faster with prognostic adjustment. This term contributes to both bias and variance of the standard error estimate. Unlike the equivalent norm that appears in the bound for the empirical process term, the norm here is not divided by $\sqrt{n}$ and so this term may be of *leading order* (slowest decaying) in the overall stochastic decomposition and thus asymptotically relevant. Since prognostic

adjustment increases the rate, the term may go from leading order to higher-order. Therefore prognostic adjustment may, in some cases, make the plug-in standard error an asymptotically linear estimator (i.e. standard error of the standard error should decrease at a $1/\sqrt{n}$ rate).

Caveats

Although we do not need additional assumptions for identifiability and thus retain unbiased estimation in all cases, all of the possible benefits described above do rely on the assumption that the historical and trial data-generating processes share a control-specific conditional mean. If this is not the case, then the amount by which the prognostic score speeds convergence of the control outcome regression will be attenuated, but not necessarily eliminated. For example, if the true outcome regression is the same as the prognostic score up to some parametric transformation that is learnable at a fast rate by a learner in our library $\mathcal{L}$, then we should still expect benefits.

Until now we have also focused on the control counterfactual mean. The influence function for the ATE is the difference of those for the two counterfactual means and consequently we can decompose the empirical process term into two terms which are $O_P(||\hat{\mu}_1 - \mu_1||n^{-1/2})$ and $O_P(||\hat{\mu}_0 - \mu_0||n^{-1/2})$ where $\mu_a = E[Y|A = a, W]$ denote the counterfactual mean functions. Therefore, the overall order of the empirical process term is dominated by whichever of these terms is lower-order. For the treatment regression to leverage the historical data, we need to assume some prognostic information can be transferred from the control to the treatment arm. For example, we might expect satisfactory information transfer from the historical data to the trial treatment arm when when there is a constant treatment effect c (i.e. $\mu_1(w) = \mu_0(w) + c$) or there is some other parametric, easily-learnable relationship between the two conditional means. Otherwise, the slower convergence rate for μ_1 dominates and we obtain less benefit from prognostic adjustment. A worst-case scenario would be when the two conditional means depend on mutually exclusive sets of covariates: if this is the case, no transfer should be possible and benefits should be limited or absent.

Our analysis shows that use of the historical sample via prognostic score adjustment produces less-variable point estimates in small samples as well as more stable and accurate estimates of standard error. Unfortunately, asymptotic gains in efficiency are not possible without further assumptions.

However, these benefits are contingent on the extent to which the covariate-outcome relationships in both treatment arms of the trial are similar to the equivalent relationship in the historical data. In particular, differences between historical and trial populations and high heterogeneity of effect may both attenuate benefits. Nonetheless, these problems can never induce bias. Therefore, relative to alternatives, prognostic adjustment of efficient estimators provides strict guarantees for type I error, but at the cost of limiting the possible benefits of using historical data.

4.4 Simulation study

Setup

This simulation study aims to demonstrate the utility of an efficient estimator with the addition of a prognostic score. We examine how our method performs in different data generating scenarios (e.g., heterogeneous vs. constant effect), across different data set sizes, and when there are distributional shifts from the historical to the trial population. The simulation study is based on the structural causal model in DGP (4.1) – (4.9). In total there are 20 observed covariates of various types and a single *unobserved* covariate.

$$W_1 \sim \text{Unif}(-2, 1) \tag{4.1}$$
$$W_2 \sim \mathcal{N}(0, 3) \tag{4.2}$$
$$W_3 \sim \text{Exp}(0.8) \tag{4.3}$$
$$W_4 \sim \Gamma(2, 1) \tag{4.4}$$
$$W_6 \dots W_20 \sim \text{Unif}(0, 1) \tag{4.5}$$
$$U \sim \text{Unif}(0, 1) \tag{4.6}$$
$$A \sim \text{Bern}(1/2) \tag{4.7}$$
$$Y^a | W, U = m_a(W, U) + \mathcal{N}(0, 2) \tag{4.8}$$
$$Y = AY^1 + (1 - A)Y^0 \tag{4.9}$$

Notice that $m_a(W, U)$ is the mean of the counterfactual conditioned on both the observed and unobserved covariates. Our observable conditional means are thus $\mu_a(W) = E[m_a(W, U)|W]$. We examine two different scenarios for the conditional outcome mean m_a. In our "heterogeneous effect" simulation:

$$m_1(W, U) = (10 \times sin(|W_1|\pi))^2 + I(U > 1.01) \times 8 + I(U > 1.55) \times 15 - 42 \tag{4.10}$$
$$m_0(W, U) = 10 \times sin(|W_1|\pi) + I(U > 1.01) \times 8 + I(U > 1.55) \times 15 \tag{4.11}$$

where the I represents the indicator function and propensity score π is written without the subscript a since the treatment probability is the same. Our "constant effect" simulation is computed as:

$$m_1(W, U) = 10 \times sin(|W_1|\pi) + I(U > 1.21) \times 20 + I(U > 1.55) \times 15 - 0.8 \tag{4.12}$$
$$m_0(W, U) = 10 \times sin(|W_1|\pi) + I(U > 1.21) \times 20 + I(U > 1.55) \times 15 \tag{4.13}$$

To begin, we use the same data generating process (DGP) for the historical and trial populations except the fact that $A = 0$ determinstically in the historical DGP. But in what follows, we loosen this assumption by changing the historical data generating distribution with varying degrees of observed and unobserved covariate shifts.

We examine several scenarios: first, we analyze the trial ($n = 250$) under the heterogeneous and constant treatment effect DGPs, where the historical sample ($\tilde{n} = 1{,}000$) is

from the same DGP (4.1) – (4.9) as the trial sample. Second, we vary the historical and trial sample sizes for the heterogeneous treatment effect simulation. To vary the historical sample sizes, we first fix the trial sample size ($n = 250$) and vary the historical sample size (with $\tilde{n} = 100$, 250, 500, 750, and 1,000). To vary the trial sample sizes, we first fix the historical sample size ($\tilde{n} = 1{,}000$) and vary the trial sample size (with $n = 100$, 250, 500, 750, and 1,000). We also vary n and $\tilde{n} = n^2$ together to demonstrate asymptotic benefits in the estimation of the standard error (as discussed in Section 4.3).

Third, we examine the effect of distributional shifts between the historical and trial populations. In these cases, we draw trial data from the DGP (4.1) – (4.9), but draw our historical data from modified versions. To simulate a "small" observable population shift we let $W_1|D = 0 \sim \text{Unif}(-5, -2)$ and to simulate a "large" observable population shift we let $W_1|D = 0 \sim \text{Unif}(-7, -4)$. To simulate a "small" *unobservable* population shift we let $U|D = 0 \sim \text{Unif}(0.5, 1.5)$ and to simulate a "large" *unobservable* population shift we let $U|D = 0 \sim \text{Unif}(1, 2)$. The shifts in the unobserved covariate induce shifts in the conditional mean relationship between the observed covariates and the outcome (see Appendix C.1 for an explicit explanation).

We consider three estimators for the trial: unadjusted (difference-in-group-means), linear regression (with Huber-White (robust) standard errors estimator HC_3[75,76,124], and targeted maximum likelihood estimation (TMLE; an efficient estimator[?]). All estimators return an effect estimate and an estimated standard error, which we use to construct Wald 95% confidence intervals and corresponding p-values. The naive unadjusted estimator cannot leverage any covariates, but both linear and TMLE estimators can. We compare and contrast results from linear and TMLE estimators both with and without the fitted prognostic score as an adjustment covariate ("fitted") to compare against Schuler et al (2021)[104]. We also consider the oracle version of the prognostic score ("oracle") for a benchmark comparison; the oracle prognostic score perfectly models the expected control outcome in the trial $E[Y|W, A = 0, D = 0]$. Unlike the fit prognostic score, the oracle version is not affected by random noise in the historical data and it is not sensitive to shifts between historical and trial populations (indeed it is not affected by the historical data at all). The oracle prognostic score only serves as a best-case comparison and is infeasible to calculate in practice.

For simplification, we include the same specifications of the discrete super learner (cross-validated ensemble algorithm) for both the prognostic model and all regressions required by our efficient estimators. Specifically, we use the discrete super learner – choosing one machine learning algorithm from a set of algorithms for each cross-fit fold via the lowest cross-validated mean squared error. The set of algorithms include the linear regression, gradient boosting with varying tree tuning specifications (xgboost)[21], and Multivariate Adaptive Regression Splines[36]. Specifications for tuning parameters are in Appendix C.2.

Results

Our results for the heterogeneous effect scenario are summarized in Table 4.1 and Table 4.2. Results for other DGPs are qualitatively similar so these are reported in Appendix C.3 along

Prognostic score	TMLE	linear	unadjusted
oracle	-0.069 (4.341)	0.034 (9.727)	-
fit	-0.064 (4.482)	0.025 (9.710)	-
none	0.009 (5.691)	0.072 (9.774)	0.153 (9.509)

Table 4.1: Empirical bias and variance for the point estimate of the trial ATE in the heterogeneous effect simulation scenario. Results in the table are formatted as "empirical bias (empirical variance)".

Table 4.2: Empirical bias and variance for the estimated standard error of the trial ATE estimate. Results in the table are formatted as "empirical bias (empirical variance)".

Prognostic score	TMLE	linear	unadjusted
oracle	0.152 (0.003)	0.138 (0.028)	-
fit	0.158 (0.005)	0.137 (0.029)	-
none	0.180 (0.055)	0.213 (0.026)	0.139 (0.011)

with additional performance metrics.

Table 4.1 illustrates the mean of the empirical bias and empirical variance of the ATE point estimate across the 200 simulations. The results demonstrate that prognostic adjustment decreases variance relative to vanilla TMLE in the realistic heterogeneous treatment effect scenario (results are similar in other scenarios). The reduction in variance results in an increase in an 11% increase in power in this case. In terms of variance reduction, fitting the prognostic score is almost as good as having the oracle in this most scenarios.

Prognostic adjustment also improves the variance of the linear estimator (corroborating Schuler et al. (2021)[104]). But overall TMLE convincingly beats the linear estimator, with or without prognostic adjustment, except for in the constant effect scenario where the two are roughly equivalent with prognostic adjustment. The matching or slightly superior performance of prognostically adjusted linear regression in the constant effect DGP is consistent with the optimality property previously discussed in Schuler et al. (2021)[104].

Importantly, the variance is not underestimated in any of our simulations meaning that the coverage was nominal (95%) for all estimators (and thus strict type I error control was attained; Appendix C.3). Including the prognostic score did not affect coverage in any case, even when the trial and historical populations were different.

Table 4.2 illustrates the mean of the empirical bias and empirical variance of the *estimated standard error* of the ATE estimate. The table corroborates the theoretical findings from Section 3, namely that the variance of the estimated variance for an efficient estimator (TMLE) is decreased by prognostic adjustment.

Using larger historical data sets increases the benefits of prognostic adjustment with efficient estimators. Figure 4.1.A shows a detailed view of this phenomenon in terms of

decrease in the average estimated standard error as the historical data set grows in size. In effect, the larger the historical data, the smaller the resulting confidence intervals tend to be in the trial (while still preserving coverage, see Appendix C.3), for the estimators leveraging an estimated prognostic score. Figure 4.1.B shows the change in estimated standard error as the *trial* size varies. This illustrates that the relative benefit of prognostic adjustment is larger in smaller trials. Here we see an 11% increase in power comparing the TMLE with versus without fitted prognostic score when $n = 250$, but an 80% increase when $n = 100$. From Figure 4.1 we again see that the TMLE with the fitted prognostic score performs almost as well as the TMLE with the oracle prognostic score when the historical sample size is increased to around 1,000.

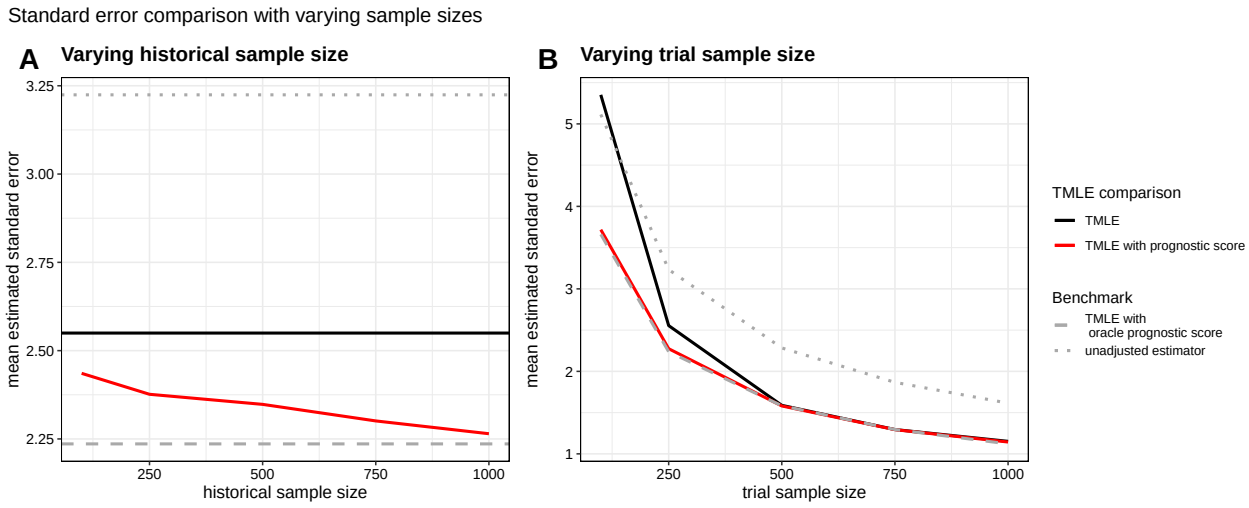


Figure 4.1: Mean estimated standard errors across estimators when historical and trial sample sizes are varied using the heterogeneous data generating process. When the historical sample size is varied (Figure 4.1.A), the trial is fixed at $n = 250$. When the trial size is varied (Figure 4.1.B), the historical sample is fixed at $\tilde{n} = 1000$.

When trial sample size n and historical sample size $\tilde{n} = n^2$ increase together, our theory predicts that the plug-in standard error may become asymptotically linear with prognostic adjustment. This is confirmed by simulation (with 1,000 Monte Carlo simulations): as we increase n with prognostic adjustment, the empirical variance of the estimated standard error times n is closer to a flat line which indicates a near-$\sqrt{n}$ rate of decay (Figure 4.2). The same is not true without prognostic adjustment: the variance of the plug-in standard error from TMLE is greater and falls more slowly.

We also observe that our method is relatively robust to both observed and unobserved distributional shifts between historical and trial populations (Figure 4.3). When the shifts are large, the prognostic score may be uninformative (most evident in Figure 4.3.B), but including it may still improve efficiency (as seen in Figure 4.3.A). We also see that a good

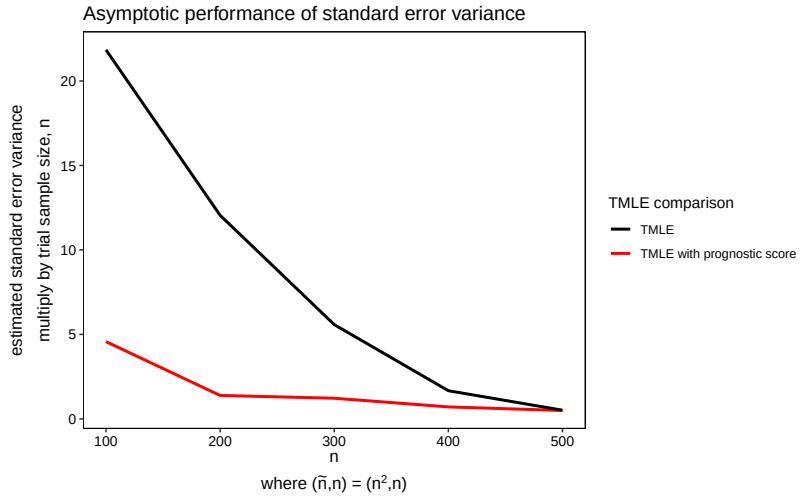


Figure 4.2: Variance of estimated standard error across estimators when historical and trial sample sizes are varied using the constant effect data generating process. The historical sample size, $\tilde{n}$ is proportional to trial size n, where $(\tilde{n}, n) = (n^2, n)$.

prognostic score (no shift in distribution) substantially reduces the variability of the estimated standard error. Variability increases with the magnitude of the covariate shift but still does not exceed that of TMLE without prognostic adjustment.

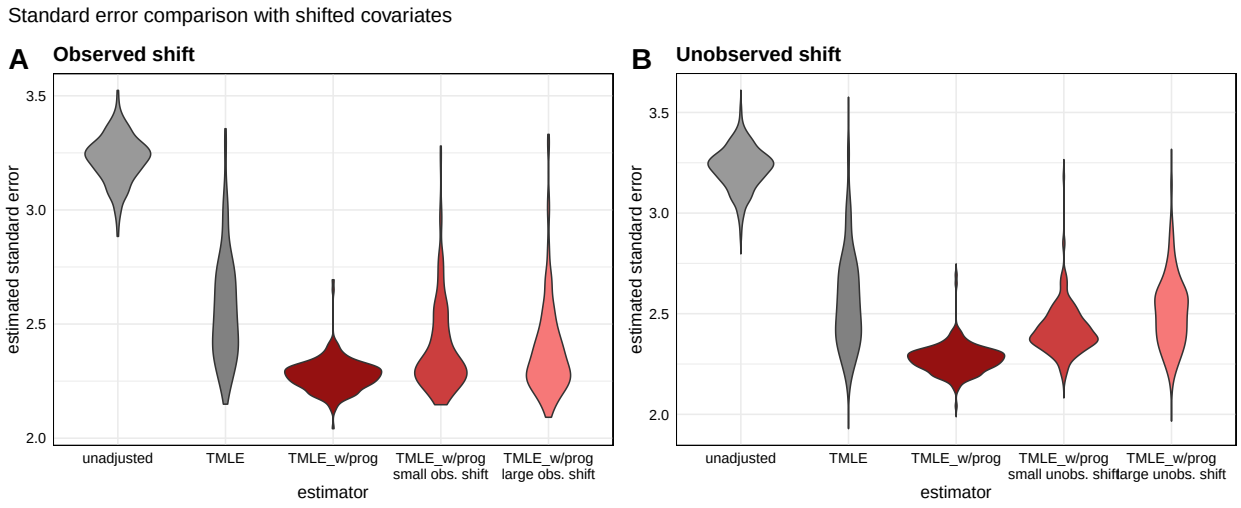


Figure 4.3: Estimated standard errors across estimators when observed (Figure 4.3.A) and unobserved shifts (Figure 4.3.B) are present in the historical sample relative to the trial sample.

4.5 Case Study

In this section, we examine the use of TMLE with prognostic covariate adjustment in RCTs involving people diagnosed with type 2 diabetes (T2D). T2D is a chronic disease with a progressive deterioration of glucose control. Glucose control is normally evaluated by long-term blood glucose level, measured by hemoglobin A1C (HbA1C). The analyses are carried out using data provided by Novo Nordisk A/S originating from 14 previously conducted RCTs within the field of diabetes, see Appendix C.4 for a full overview of the trials. Our use of data for this study is approved by Novo Nordisk A/S.

We reanalyse the phase IIIb clinical trial called NN9068-4229, where the trial population consisted of insulin naive people with T2D[25]. The participants of this trial were inadequately controlled on treatment with SGLT2i, a type of oral anti-diabetic treatment (OAD). Inadequately controlled was defined as having a HbA1C of 7.0-11.0% (both inclusive). The aim of the trial was to compare glycemic control of insulin IDegLira versus insulin IGlar as add-on therapy to SGLT2i in people with T2D. The trial was a 26-week, 1:1 randomized, active-controlled, open label, treat-to-target trial with 420 enrolled participants. One participant was excluded due to non-exposure to trial product, yielding $n = 419$. The efficacy of IDegLira was measured by the difference in change from baseline HbA1C to landmark visit week 26. Our corresponding historical sample came from previously conducted RCTs with a study population also consisting of insulin naive people with T2D, who were inadequately controlled on their current OADs. A total of $\tilde{n} = 3311$ participants all receiving insulin IGlar, were enrolled in the historical sample.

For the trial reanalysis in our study, we included patient measures of their demographic background, laboratory measures, concomitant medication, and vital signs. The treatment indicator where only used in the NN9068-4229 trial. For details on the specific measurements, covariate distributions, and imputation of missing covariates see Appendix C.5, C.6, and C.7. For the continuous covariates we see that the mean and standard deviation are not particularly different between the historical and new trial sample, meaning that both resemble a T2D population with uncontrolled glycemic control. Furthermore we see that the range of continuous covariates for the new trial sample are contained in the range of the historical sample. This indicates that the trial population is largely similar to the historical population, at least in terms of observable covariates. For the categorical covariates the distributions vary between the historical and new trial sample. However, all the categories in the new trial sample are present in the historical sample.

A linear estimator with baseline HbA1C, region and pre-trial OADs as adjustment covariates was used in the original analysis of the primary endpoint in the NN9068-4229 trial. In this analysis, an average treatment effect estimate of -0.340 (95% confidence interval [-0.480;-0.200]). In this reanalysis, we report the result of five estimators: unadjusted, linear regression (adjusting for all available covariates), linear regression with a prognostic score, TMLE, and TMLE with a prognostic score. For this application, we expanded the library of the super learner for a more comprehensive set of machine learning models than the simulation, including random forest[17], k-nearest neighbor, and a more comprehensive set of

tuning parameters for the xgboost model in addition to the previously specified library, see Appendix C.2. Separately, we obtained the correlation of the fitted prognostic score against the trial outcome. The prognostic score's correlation with the outcome is 0.752 with control subjects and 0.622 with treated subjects, indicating that adjustment for the score should result in an improvement over unadjusted estimation [104].

Table 4.3: Estimates for average treatment effect and with 95% confidence levels of change in hemoglobin A1C from baseline to week 26 for insulin IDegLira versus insulin IGlar as add-on therapy to SGLT2i in people with type 2 diabetes.

Prognostic score	TMLE	linear	unadjusted
with	-0.351 (s.e. 0.145)	-0.355 (s.e. 0.157)	-
without	-0.369 (s.e. 0.150)	-0.355 (s.e. 0.157)	-0.248 (0.192)

This is a reanalyses of the NN9068-4229 trial using five different estimators where 1:1 randomization was performed. The total sample size is $n = 419$.

From Table 4.3, we see that the smallest confidence interval is obtained using TMLE with prognostic score. All methods obtain similar point estimates except from the unadjusted estimator. Notice that the linear estimator with or without a prognostic score yields the same results, since the prognostic model is a linear model in this case (chosen as the model that yielded lowest MSE from 20-fold cross-validation within the discrete super learner).

Table 4.4: Estimates for average treatment effect and with 95% confidence levels of change in hemoglobin A1C from baseline to week 26 for insulin IDegLira versus insulin IGlar as add-on therapy to SGLT2i in people with type 2 diabetes.

Prognostic score	TMLE	linear	unadjusted
with	-0.519 (s.e. 0.307)	-0.544 (s.e. 0.438)	-
without	-0.582 (s.e. 0.349)	-0.544 (s.e. 0.438)	-0.344 (0.399)

This is a reanalyses of the NN9068-4229 trial using five different estimators where 50 participants from the control and treatment group, respectively, were chosen at random yielding a total sample size of $n = 100$. The random selection is done 10 times and the reported numbers are the average of the point estimates and standard error.

As illustrated by the simulation study and the asymptotic analysis in Sections 4.3 and 4.3, the relative benefit of prognostic adjustment is larger in smaller trials. To examine this result, we sub-sampled from the the NN9068-4229 trial but reanalyzed with selecting 50 participants randomly from each group, resulting in $n = 100$. This random selection of 50 participants from each group is repeated 10 times and averaged to compute the point estimate and standard error. The average correlation of the prognostic score with the outcome was 0.790 with control subjects and 0.656 with treated subjects. We see an relatively larger

reduction in the standard error estimate using TMLE with prognostic covariate adjustment compared to TMLE without in the reanalysis (Table 4.4).

4.6 Discussion

In this study we demonstrate the utility of incorporating historical data via a prognostic score in an efficient estimator while maintaining strict type I error control. Using the prognostic score via covariate adjustment overall improves the performance of the efficient estimator by decreasing the standard error and improving its estimation. This method is most useful in randomized trials with small sample sizes. Our proposed method is shown to be robust against bias even when the historical sample is drawn from a different population.

Prognostic adjustment requires no assumptions to continue to guarantee unbiased causal effect estimates. However, this comes with a trade-off: without introducing the risk of bias, there is a limit on how much power can be gained and in what scenarios. For example, the method of Li et al. (2021) (which imposes an additional assumption) can asymptotically benefit from the addition of historical data, whereas our method can only provide gains in small samples[70]. However, these gains are *most important* precisely in small samples because estimated effects are likely to be of borderline significance, whereas effects are more likely to be clear in very large samples regardless of the estimator used.

Besides being assumption-free, our method has other practical advantages relative to data fusion approaches. For one, we do not require a single, well-defined treatment in the historical data. Moreover, we do not require an exact overlap of the covariates measured in the historical and trial data sets. It is also easy to utilize multiple historical data sets: if they are believed to be drawn from substantially different populations, separate prognostic scores can be built from each of them and included as covariates in the trial analysis. As long as one of these scores is a good approximation of the outcome-covariate relationship in one or more arms of the trial, there will be added benefits to power.

Prognostic adjustment with efficient estimators can also be used with pre-built or public prognostic models: the analyst does not need direct access to the historical data if they can query a model for predictions. This is helpful in cases where data is "federated" and cannot move (e.g. when privacy must be protected or data has commercial value).

The theory we developed to explain the benefits of prognostic adjustment in the context of efficient estimation for trials is easily generalizable to estimation of any kind of pathwise differentiable parameter augmented with transfer learning from an auxiliary dataset. The specific breakdown of different terms may differ but the overall intuition should be the same: transfer learning may accelerate the disappearance of higher-order terms that depend on the error rates of regression estimates.

Our approach is closely related to the transfer learning literature in machine learning. In transfer learning, the goal is to use a (large) "source" dataset to improve prediction for a "target" population for which we have only minimal training data[117,123,132]. In this work we use a particular method of "transfer" (adjusting for the source/historical model prediction)

to improve the target (trial) predictions, which drives variance reduction. It should also be possible to leverage other more direct forms of model transfer for the outcome regression, such as pre-training a deep learning model on the historical data and then fine-tuning using the trial data.

Lastly, since we use efficient estimators, we can leverage the results of Schuler (2021) to prospectively calculate power with prognostic adjustment [102]. In fact, we suspect the methods of power calculation described in that work would improve in accuracy with prognostic adjustment since the outcome regressions are "jump-started" with the prognostic score. Verification of this fact and empirical demonstration will be left to future work.

Bibliography

[1] A. Abadie, M. M. Chingos, and M. R. West. Endogenous stratification in randomized experiments. Review of Economics and Statistics, 100(4):567–580, 2018.

[2] A. Agarwal, A. M. Kenney, Y. S. Tan, T. M. Tang, and B. Yu. Mdi+: A flexible random forest-based feature importance framework. arXiv preprint arXiv:2307.01932, 2023.

[3] A. Ahmed, A. Husain, T. E. Love, G. Gambassi, L. J. Dell'Italia, G. S. Francis, M. Gheorghiade, R. M. Allman, S. Meleth, and R. C. Bourge. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. European heart journal, 27(12):1431–1439, 2006.

[4] R. C. Aikens and M. Baiocchi. Assignment-control plots: A visual companion for causal inference study design. The American Statistician, pages 1–13, 2022.

[5] R. C. Aikens, D. Greaves, and M. Baiocchi. A pilot design for observational studies: Using abundant data thoughtfully. Statistics in Medicine, 39(30):4821–4840, 2020.

[6] A. Akbari, A. Fathabadi, M. Razmi, A. Zarifian, M. Amiri, A. Ghodsi, and E. V. Moradi. Characteristics, risk factors, and outcomes associated with readmission in covid-19 patients: A systematic review and meta-analysis. The American journal of emergency medicine, 52:166–173, 2022.

[7] American College of Obstetricians and Gynecologists. Acog practice bulletin no. 190: gestational diabetes mellitus. Obstet Gynecol, 131(2):e49–e64, 2018.

[8] S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4):597–623, 2018.

[9] J. L. Atkins, J. A. Masoli, J. Delgado, L. C. Pilling, C.-L. Kuo, G. A. Kuchel, and D. Melzer. Preexisting comorbidities predicting covid-19 and mortality in the uk biobank community cohort. The Journals of Gerontology: Series A, 75(11):2224–2230, 2020.

[10] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. Proc. Natl. Acad. Sci. U. S. A., 113(27):7345–7352, July 2016.

[11] O. Y. Bello-Chavolla, J. P. Bahena-López, N. E. Antonio-Villa, A. Vargas-Vázquez, A. González-Díaz, A. Márquez-Salinas, C. A. Fermín-Martínez, J. J. Naveja, and C. A. Aguilar-Salinas. Predicting mortality due to sars-cov-2: a mechanistic score relating obesity and diabetes to covid-19 outcomes in mexico. The Journal of Clinical Endocrinology & Metabolism, 105(8):2752–2761, 2020.

[12] E. Ben-Michael, A. Feller, D. A. Hirshberg, and J. R. Zubizarreta. The balancing act in causal inference. arXiv preprint arXiv:2110.14831, 2021.

[13] E. Ben-Michael, A. Feller, and J. Rothstein. Varying impacts of letters of recommendation on college admissions: Approximate balancing weights for subgroup effects in observational studies. arXiv preprint arXiv:2008.04394, 2021.

[14] M. Bennett, J. P. Vielma, and J. R. Zubizarreta. Building representative matched samples with multi-valued treatments in large observational studies. Journal of Computational and Graphical Statistics, 29(4):744–757, 2020.

[15] C. Bentley, S. Cressman, K. van der Hoek, K. Arts, J. Dancey, and S. Peacock. Conducting clinical trials—costs, impacts, and the value of clinical trials networks: A scoping review. Clin. Trials, 16(2):183–193, Apr. 2019.

[16] C. Bicalho, A. Bouyamourn, and T. Dunning. Conditional balance tests: Increasing sensitivity and specificity with prognostic covariates. arXiv preprint arXiv:2205.10478, 2022.

[17] L. Breiman. Random forests. Machine learning, 45:5–32, 2001.

[18] J. M. Brooks and R. L. Ohsfeldt. Squeezing the balloon: propensity scores and unmeasured covariate balance. Health services research, 48(4):1487–1507, 2013.

[19] W. C. Castillo, K. Boggess, T. Stürmer, M. A. Brookhart, D. K. Benjamin Jr, and M. J. Funk. Trends in glyburide compared with insulin use for gestational diabetes treatment in the united states, 2000-2011. Obstetrics and gynecology, 123(6):1177, 2014.

[20] A. Chakrabortty, G. Dai, and E. T. Tchetgen. A general framework for treatment effect estimation in semi-supervised and high dimensional settings. arXiv preprint arXiv:2201.00468, 2022.

[21] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

[22] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. Econom. J., 21(1):C1–C68, Jan. 2018.

[23] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. Project Euclid, 2010.

[24] C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82 (1):39–67, 2020.

[25] ClinicalTrials.gov. A clinical trial comparing glycaemic control and safety of insulin degludec/liraglutide (ideglira) versus insulin glargine (iglar) as add-on therapy to sglt2i in subjects with type 2 diabetes mellitus (dual tm ix), 2020. URL `https://clinicaltrials.gov/study/NCT02773368?cond=DUALTM%20IX%20-%20Add-on%20to%20SGLT2i&rank=1`. Accessed on 2023-9-25.

[26] B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. Statistical science, 39(1):165–191, 2024.

[27] D. Cox. On a generalization of a result of wg cochran. Biometrika, 94(3):755–759, 2007.

[28] L. E. Dang, J. M. Tarp, T. J. Abrahamsen, K. Kvist, J. B. Buse, M. Petersen, and M. van der Laan. A cross-validated targeted maximum likelihood estimator for data-adaptive experiment selection applied to the augmentation of rct control arms with external data. arXiv preprint arXiv:2210.05802, 2022.

[29] V. de variantes del virus SARS-CoV-2. Vigilancia de variantes del virus sars-cov-2, 2022. URL `https://salud.conacyt.mx/coronavirus/variantes/`. Accessed on 2022-7-29.

[30] I. Degtiar and S. Rose. A review of generalizability and transportability. Annu. Rev. Stat. Appl., 10(1):501–524, Mar. 2023.

[31] D. K. Dey, S. K. Ghosh, and B. K. Mallick. Generalized linear models: A Bayesian perspective. CRC Press, 2000.

[32] I. Diaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. Biostatistics, 21(2):353–358, Apr. 2020.

[33] P. Ding, T. VanderWeele, and J. M. Robins. Instrumental variables as bias amplifiers with general outcome and confounding. Biometrika, 104(2):291–302, 2017.

[34] J. Escobedo-de la Peña, R. A. Rascón-Pacheco, I. de Jesús Ascencio-Montiel, E. González-Figueroa, J. E. Fernández-Gárate, O. S. Medina-Gómez, P. Borja-Bustamante, J. A. Santillán-Oropeza, and V. H. Borja-Aburto. Hypertension, diabetes and obesity, major risk factors for death in patients with covid-19 in mexico. Archives of medical research, 52(4):443–449, 2021.

[35] FDA. Adjusting for covariates in randomized clinical trials for drugs and biological products, 2023.

[36] J. H. Friedman. Multivariate adaptive regression splines. The annals of statistics, 19 (1):1–67, 1991.

[37] E. Ge, Y. Li, S. Wu, E. Candido, and X. Wei. Association of pre-existing comorbidities with mortality and disease severity among 167,500 individuals with covid-19 in canada: A population-based cohort study. PLoS One, 16(10):e0258154, 2021.

[38] T. V. Giannouchos, R. A. Sussman, J. M. Mier Odriozola, K. Poulas, and K. Farsalinos. Characteristics and risk factors for covid-19 diagnosis and adverse outcomes in mexico: an analysis of 89,756 laboratory–confirmed covid-19 cases. Medrxiv, pages 2020–06, 2020.

[39] A. Gimeno-Miguel, K. Bliek-Bueno, B. Poblador-Plou, J. Carmona-Pírez, A. Poncel-Falcó, F. González-Rubio, I. Ioakeim-Skoufa, V. Pico-Soler, M. Aza-Pascual-Salcedo, A. Prados-Torres, et al. Chronic diseases associated with increased likelihood of hospitalization and mortality in 68,913 covid-19 confirmed cases in spain: A population-based cohort study. PloS one, 16(11):e0259822, 2021.

[40] G. V. Glass. Primary, secondary, and meta-analysis of research. Educational researcher, 5(10):3–8, 1976.

[41] R. Glennerster. Chapter 5 - the practicalities of running randomized evaluations: Partnerships, measurement, ethics, and transparency. In A. V. Banerjee and E. Duflo, editors, Handbook of Economic Field Experiments, volume 1, pages 175–243. North-Holland, Jan. 2017.

[42] A. N. Glynn and K. M. Quinn. An introduction to the augmented inverse propensity weighted estimator. Polit. Anal., 18(1):36–56, 2010.

[43] N. Greifer. cobalt: Covariate Balance Tables and Plots, 2021. URL `https://CRAN.R-project.org/package=cobalt`. R package version 4.3.1.

[44] N. Greifer and E. A. Stuart. Choosing the estimand when matching or weighting in observational studies. arXiv preprint arXiv:2106.10577, 2021.

[45] K. Guo and D. Rothenhäusler. On the statistical role of inexact matching in observational studies. Biometrika, 110(3):631–644, 2023.

[46] J. P. Gutierrez and S. M. Bertozzi. Non-communicable diseases and inequalities increase risk of death among covid-19 patients in mexico. PloS one, 15(10):e0240394, 2020.

[47] C. Hans. Elastic net regression modeling with the orthant normal prior. Journal of the American Statistical Association, 106(496):1383–1393, 2011.

[48] B. B. Hansen. Full matching in an observational study of coaching for the sat. Journal of the American Statistical Association, 99(467):609–618, 2004.

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

[50] B. B. Hansen and J. Bowers. Covariate balance in simple, stratified and clustered comparative studies. Statistical Science, pages 219–236, 2008.

[51] B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. Journal of computational and Graphical Statistics, 15(3):609–627, 2006.

[52] L. V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. journal of Educational Statistics, 6(2):107–128, 1981.

[53] S. Heng, H. Kang, D. S. Small, and C. B. Fogarty. Increasing power for observational studies of aberrant response: An adaptive approach. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(3):482–504, 2021.

[54] D. R. Hernández-Galdamez, M. Á. González-Block, D. K. Romo-Dueñas, R. Lima-Morales, I. A. Hernández-Vicente, M. Lumbreras-Guzmán, and P. Mendez-Hernandez. Increased risk of hospitalization and death in patients with covid-19 and pre-existing noncommunicable diseases and modifiable risk factors in mexico. Archives of medical research, 51(7):683–689, 2020.

[55] J. L. Hill. Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat., 20(1):217–240, Jan. 2011.

[56] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

[57] B. Holzhauer and E. T. Adewuyi. "super-covariates": Using predicted control group outcome as a covariate in randomized clinical trials. Pharmaceutical Statistics, 22(6): 1062–1075, 2023.

[58] J. Hopkins. University of medicine. coronavirus resource center. Data Stream, 2020.

[59] S. R. Howard and S. D. Pimentel. The uniform general signed rank test and its design sensitivity. Biometrika, 108(2):381–396, 2021.

[60] M. Huang and S. D. Pimentel. Variance-based sensitivity analysis for weighting estimators result in more informative bounds. arXiv preprint arXiv:2208.01691, 2022.

[61] M. Huang, N. Egami, E. Hartman, and L. Miratrix. Leveraging population outcomes to improve the generalization of experimental results: Application to the jtpa study. The Annals of Applied Statistics, 17(3):2139–2164, 2023.

[62] M. Huang, D. Soriano, and S. D. Pimentel. Design sensitivity and its implications for weighted observational studies. arXiv preprint arXiv:2307.00093, 2023.

[63] A. Juárez-Flores, I. J. Ascencio-Montiel, J. P. Gutiérrez, S. M. Bertozzi, V. H. Borja-Aburto, and G. Olaiz. Covid-19 in the mexican social security institute (imss) population. prevalent symptoms. medRxiv, pages 2022–04, 2022.

[64] S. Kastora, M. Patel, B. Carter, M. Delibegovic, and P. K. Myint. Impact of diabetes on covid-19 mortality and hospital outcomes from a global perspective: An umbrella systematic review and meta-analysis. Endocrinology, diabetes & metabolism, 5(3): e00338, 2022.

[65] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint arXiv:2203.06469, 2022.

[66] E. LeDell, M. Petersen, and M. van der Laan. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. Electronic journal of statistics, 9(1):1583, 2015.

[67] D. Lee, S. Yang, L. Dong, X. Wang, D. Zeng, and J. Cai. Improving trial generalizability using observational studies. Biometrics, Dec. 2021.

[68] F. Li, L. E. Thomas, and F. Li. Addressing extreme propensity scores via the overlap weights. American journal of epidemiology, 188(1):250–257, 2019.

[69] F. Li, P. Ding, and F. Mealli. Bayesian causal inference: a critical review. Philos. Trans. A Math. Phys. Eng. Sci., 381(2247):20220153, May 2023.

[70] X. Li, W. Miao, F. Lu, and X.-H. Zhou. Improving efficiency of inference in clinical trials with external control data. Biometrics, Oct. 2021.

[71] L. D. Liao and S. D. Pimentel. jointvip: Prioritizing variables in observational study design with joint variable importance plot in r. arXiv preprint arXiv:2302.10367, 2023.

[72] L. D. Liao, A. E. Hubbard, J. P. Gutierrez, A. Juárez-Flores, K. Kikkawa, R. Gupta, Y. Yarmolich, I. de Jesús Ascencio-Montiel, and S. M. Bertozzi. Who is most at risk of dying if infected with sars-cov-2? a mortality risk factor analysis using machine learning of patients with covid-19 over time: A large population-based cohort study in mexico. BMJ open, 13(9):e072436, 2023.

[73] L. D. Liao, Y. Zhu, A. L. Ngo, R. F. Chehab, and S. D. Pimentel. Prioritizing variables for observational study design using the joint variable importance plot. The American Statistician, pages 1–9, 2024.

[74] H. Liu. Generalized additive model. Department of Mathematics and Statistics University of Minnesota Duluth: Duluth, MN, USA, 55812, 2008.

[75] J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. The American Statistician, 54(3):217–224, 2000. doi: http://dx.doi.org/10.1080/00031305.2000.10474549.

[76] J. G. MacKinnon and H. White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics, 29(3): 305–325, 1985. doi: https://doi.org/10.1016/0304-4076(85)90158-7.

[77] M. U. Martínez-Martínez, D. Alpízar-Rodríguez, R. Flores-Ramírez, D. P. Portales-Pérez, R. E. Soria-Guerra, F. Pérez-Vázquez, and F. Martinez-Gutierrez. An analysis covid-19 in mexico: a prediction of severity. Journal of general internal medicine, pages 1–8, 2022.

[78] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2):153–157, 1947.

[79] L. W. Miratrix, J. S. Sekhon, A. G. Theodoridis, and L. F. Campos. Worth weighting? how to think about and use weights in survey experiments. Political Analysis, 26(3): 275–291, 2018.

[80] K. L. Moore and M. J. van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. Stat. Med., 28(1):39–64, Jan. 2009.

[81] D. E. Noyola, N. Hermosillo-Arredondo, C. Ramírez-Juárez, and A. Werge-Sánchez. Association between obesity and diabetes prevalence and covid-19 mortality in mexico: an ecological study. The Journal of Infection in Developing Countries, 15(10):1396–1403, 2021.

[82] G. M. Parra-Bracamonte, N. Lopez-Villalobos, and F. E. Parra-Bracamonte. Clinical characteristics and risk factors for mortality of patients with covid-19 in a large data set from mexico. Annals of epidemiology, 52:93–98, 2020.

[83] K. Pearson. Vii. note on regression and inheritance in the case of two parents. proceedings of the royal society of London, 58(347-352):240–242, 1895.

[84] M. L. Petersen and M. J. van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology, 25(3):418–426, May 2014.

[85] S. D. Pimentel and Y. Huang. Covariate-adaptive randomization inference in matched designs. arXiv preprint arXiv:2207.05019, 2023.

[86] S. D. Pimentel, R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. Journal of the American Statistical Association, 110(510):515–527, 2015.

[87] S. D. Pimentel, D. S. Small, and P. R. Rosenbaum. Constructed second control groups and attenuation of unmeasured biases. Journal of the American Statistical Association, 111(515):1157–1167, 2016.

[88] E. C. Polley and M. J. van der Laan. Super learner in prediction. U.C. Berkeley Division of Biostatistics Working Paper Series, 2010.

[89] K. E. Porter, S. Gruber, M. J. Van Der Laan, and J. S. Sekhon. The relative performance of targeted maximum likelihood estimators. The international journal of biostatistics, 7(1):000010220215574679 1308, 2011.

[90] P. R. Rosenbaum. Model-based direct adjustment. Journal of the American statistical Association, 82(398):387–394, 1987.

[91] P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. Statistical Science, 17(3):286–327, 2002.

[92] P. R. Rosenbaum. Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. The American Statistician, 59(2):147–152, 2005.

[93] P. R. Rosenbaum. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics, 63(2):456–464, 2007.

[94] P. R. Rosenbaum. Design of observational studies, volume 10. Springer, 2010.

[95] P. R. Rosenbaum. Design sensitivity and efficiency in observational studies. Journal of the American Statistical Association, 105(490):692–702, 2010.

[96] P. R. Rosenbaum. Optimal matching of an optimally chosen subset in observational studies. Journal of Computational and Graphical Statistics, 21(1):57–71, 2012.

[97] P. R. Rosenbaum and D. B. Rubin. The bias due to incomplete matching. Biometrics, pages 103–116, 1985.

[98] M. Rosenblum and M. J. van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. The international journal of biostatistics, 6(1):Article 13, Apr. 2010.

[99] C. Rothe. Flexible covariate adjustments in randomized experiments, 2018. URL `http://www.christophrothe.net/papers/fca_apr2020.pdf`. Accessed: 2023-5-2.

[100] A. C. Sales, B. B. Hansen, and B. Rowan. Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. Journal of Educational and Behavioral Statistics, 43(1):3–31, 2018.

[101] W. Sawadogo, M. Tsegaye, A. Gizaw, and T. Adera. Overweight and obesity as risk factors for covid-19-associated hospitalisations and death: systematic review and meta-analysis. BMJ nutrition, prevention & health, 5(1):10, 2022.

[102] A. Schuler. Designing efficient randomized trials: power and sample size calculation when using semiparametric efficient estimators. The international journal of biostatistics, 18(1):151–171, Aug. 2021.

[103] A. Schuler and M. van der Laan. Introduction to modern causal inference. https://alejandroschuler.github.io/mci, 2022. Accessed: 2023-9-12.

[104] A. Schuler, D. Walsh, D. Hall, J. Walsh, C. Fisher, Critical Path for Alzheimer's Disease, Alzheimer's Disease Neuroimaging Initiative, and Alzheimer's Disease Cooperative Study. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. The international journal of biostatistics, Dec. 2021.

[105] L. Semenzato, J. Botton, J. Drouin, F. Cuenot, R. Dray-Spira, A. Weill, and M. Zureik. Chronic diseases, health conditions and risk of covid-19-related hospitalization and in-hospital mortality during the first wave of the epidemic in france: a cohort study of 66 million people. The Lancet Regional Health–Europe, 8, 2021.

[106] W. Shang, Y. Wang, J. Yuan, Z. Guo, J. Liu, and M. Liu. Global excess mortality during covid-19 pandemic: A systematic review and meta-analysis. Vaccines, 10(10): 1702, 2022.

[107] X. Shi, Z. Pan, and W. Miao. Data integration in causal inference. Wiley Interdisciplinary Reviews: Computational Statistics, 15(1), Jan. 2023.

[108] M. Singer. Deadly companions: Covid-19 and diabetes in mexico. Medical Anthropology, 39(8):660–665, 2020.

[109] R. d. C. M. Soares, L. R. Mattos, and L. M. Raposo. Risk factors for hospitalization and mortality due to covid-19 in espírito santo state, brazil. The American journal of tropical medicine and hygiene, 103(3):1184, 2020.

[110] D. Soriano, E. Ben-Michael, P. J. Bickel, A. Feller, and S. D. Pimentel. Interpretable sensitivity analysis for balancing weights. arXiv preprint arXiv:2102.13218, 2021.

[111] D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 2012.

[112] D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 2012.

[113] E. A. Stuart. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1): 1, 2010.

[114] E. A. Stuart, G. King, K. Imai, and D. Ho. Matchit: nonparametric preprocessing for parametric causal inference. Journal of statistical software, 2011.

[115] R. Temple and S. S. Ellenberg. Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: ethical and scientific issues. Ann. Intern. Med., 133(6):455–463, Sept. 2000.

[116] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288, 1996.

[117] L. Torrey and J. Shavlik. Transfer learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pages 242–264. IGI Global, 2010.

[118] M. J. Van der Laan and S. Rose. Targeted learning in data science. Springer, 2018.

[119] M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. The international journal of biostatistics, 2(1), 2006.

[120] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1):Article25, Sept. 2007.

[121] M. J. Van der Laan, S. Rose, et al. Targeted learning: causal inference for observational and experimental data, volume 4. Springer, 2011.

[122] K. Van Lancker, F. Bretz, and O. Dukes. The use of covariate adjustment in randomized controlled trials: An overview. arXiv preprint arXiv:2306.05823, 2023.

[123] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. Journal of Big Data, 3(1):1–40, May 2016.

[124] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48(4):817–838, 1980. doi: doi:https://doi.org/10.2307/1912934.

[125] S. Wollenstein-Betech, C. G. Cassandras, and I. C. Paschalidis. Personalized predictive models for symptomatic covid-19 patients using basic preconditions: hospitalizations, mortality, and the need for an icu or ventilator. International Journal of Medical Informatics, 142:104258, 2020.

[126] S. Wollenstein-Betech, A. A. Silva, J. L. Fleck, C. G. Cassandras, and I. C. Paschalidis. Physiological and socioeconomic characteristics predict covid-19 mortality and resource utilization in brazil. PloS one, 15(10):e0240346, 2020.

[127] P. Wu, S. Luo, and Z. Geng. On the comparative analysis of average treatment effects estimation via data combination. arXiv preprint arXiv:2311.00528, 2023.

[128] D. Yang, D. S. Small, J. H. Silber, and P. R. Rosenbaum. Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics, 68(2):628–636, 2012.

[129] Y. Zhang and Q. Zhao. What is a randomization test? arXiv preprint arXiv:2203.10980, 2022.

[130] A. Zhao and P. Ding. Covariate-adjusted fisher randomization tests for the average treatment effect. Journal of Econometrics, 225(2):278–294, 2021.

[131] H. Zhao and S. Yang. Outcome-adjusted balance measure for generalized propensity score model selection. Journal of Statistical Planning and Inference, 221:188–200, 2022.

[132] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. Proc. IEEE, 109(1):43–76, Jan. 2021.

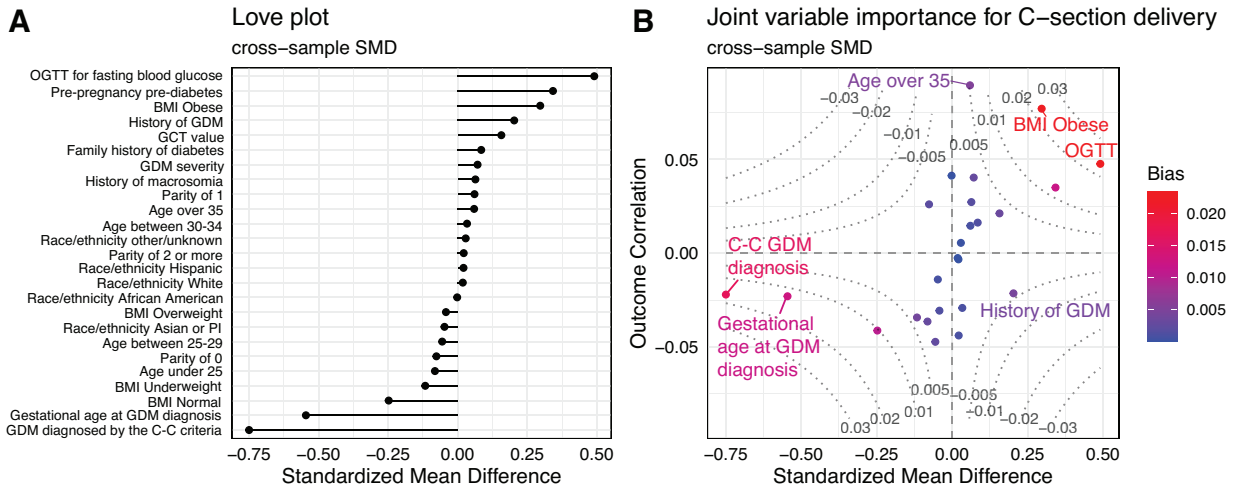
[133] J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511):910–922, 2015.

[134] J. R. Zubizarreta, R. D. Paredes, and P. R. Rosenbaum. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. The Annals of Applied Statistics, 8(1):204–231, 2014.

[135] J. R. Zubizarreta, E. A. Stuart, D. S. Small, and P. R. Rosenbaum. Handbook of matching and weighting adjustments for causal inference. CRC Press, 2023.

Appendix A Supplementary Material for Chapter 2

A.1 Comparison between the Love plot and the joint treatment-outcome variable importance plot with signed measures



Appendix Figure A.1: Comparison between the Love plot and the joint treatment-outcome variable importance plot with signed measures.

A.2 Unadjusted bias as normalized expected bias in a finite population framework

We now show that the change in the expected value of the finite population bias in equation (10) associated with matching on one additional covariate X is related to the unadjusted bias formula (8) when outcomes are drawn from a particular distribution. Assume the finite population framework of Section 2.3 but suppose that in addition all $2K$ potential outcomes under control in the study come from the following model:

$$Y(0) = \beta_0 X + \epsilon. \tag{14}$$

Note that this assumption does not change our strategy for inference, which will condition on the realized values of $Y(1)$ and $Y(0)$; rather, we invoke it only at the design stage (i.e. prior to observing or using study outcomes from the analysis sample) to help select a match. Considering hypothetical outcome distributions in this manner in this way is common in the matching literature, even though outcomes are not considered random for purposes of inference[53,59,95].

Consider the expected value of expression (10) over all realizations of $Y(0)s$ sampled from model (14). We rewrite the resulting expression in terms of an expectation over all possible treatment assignments Z_{ki} within matched pairs, conditional on the pairs themselves and the covariates. The shorthand notations E_Y and E_Z will be used to denote expectations over these two distinct types of random variation.

$$\begin{aligned}
&E_Y\left[\frac{1}{K}\sum_{k=1}^{K}[Y_{k1}(0)-Y_{k2}(0)](p_{k1}-p_{k2})\,\middle|\, X_{11},\ldots,X_{K2}\right]\\
&=\frac{1}{K}\sum_{k=1}^{K}E_Y[Y_{k1}(0)-Y_{k2}(0)\mid X_{11},\ldots,X_{K2}](p_{k1}-p_{k2})\\
&=\beta_0\cdot\left[\frac{1}{K}\sum_{k=1}^{K}(X_{k1}-X_{k2})(p_{k1}-p_{k2})\right]\\
&=\beta_0\cdot E_Z\left[\frac{1}{K}\sum_{k=1}^{K}(X_{k1}-X_{k2})(Z_{k1}-Z_{k2})\right]\\
&=\beta_0\cdot E_Z\left[\bar{X}_{1,matched}-\bar{X}_{0,matched}\right] \qquad (15)
\end{aligned}$$

where $\bar{X}_{z,matched}$ indicates the sample mean of covariate values for matched individuals with observed treatment z.

Now compare this quantity to the unadjusted bias formula (8), which we reprint here for easier comparison:

$$\frac{\Delta_j\beta_j}{S_{Y_{pilot}}}=r_{X_{j,pilot},Y_{pilot}}\frac{(\bar{X}_{j1,analysis}-\bar{X}_{j0,analysis})}{S_{X_{j,pilot}}}$$

The β_j and Δ_j terms are sample quantities, while the terms in expression (15) are parameters. However, under model (14) β_0 is the expected value of β_j. The link between $\Delta_j=\frac{(\bar{X}_{j1,analysis}-\bar{X}_{j0,analysis})}{S_{X_{j,pilot}}}$ and $E\left[\bar{X}_{1,matched}-\bar{X}_{0,matched}\right]$ is not as immediate; the scaling factor $S_{X_{j,pilot}}$ is present in the first term but not the second, and they differ in whether covariate imbalance is measured before or after matching. However, to construct a design-stage diagnostic, it is reasonable to view $\bar{X}_{j1,analysis}-\bar{X}_{j0,analysis}$ as an approximation to $\bar{X}_{1,matched}-\bar{X}_{0,matched}$ in the case where we ignore variable X when matching. Under this interpretation, we may view unadjusted bias as a normalized estimate of the bias incurred by ignoring variable X when matching as opposed to matching exactly on it.

A.3 Simulation study

A.3.1 Set up

We conduct a simulation to assess jointVIP's ability to reduce bias empirically. Our data generating process is structured with 5 confounders Xs (contribute to both treatment and

outcome regressions), 30 variables Ws contributed to treatment only, 3 variables Vs contributed to outcome only, and 30 variables Rs contributing to neither treatment or outcome dimensions. I denotes the indicator function.

$$
\begin{aligned}
X_i &\sim I(Unif(0,1) > 0.5) \text{ where } i \in \{1, ..., 5\} \\
W_j &\sim I((Unif(0,1) - 0.4) > 0.5) \text{ where } j \in \{1, ..., 30\} \\
V_k &\sim I((Unif(0,1)) > 0.5) \text{ where } k \in \{1, 2, 3\} \\
R_l &\sim I((Unif(0,1)) > 0.5) \text{ where } l \in \{1, ..., 30\}
\end{aligned}
$$

This yields 68 observed covariates in total. The treatment and outcome regressions are specified linearly with a constant treatment effect of **0.5**.

$$
\begin{aligned}
Z &\sim Binom(1, \frac{1}{1 + e^{-((0.2 * \sum_{i=1}^{5} X_i) + (0.5 * \sum_{j=1}^{30} (-1)^{(j)} * W_j) - 3)}}) \\
Y &= 3 * \sum_{k=1}^{3} V_k - 2 * \sum_{i=1}^{5} X_i + 0.5 * Z + \epsilon
\end{aligned}
$$

Here ϵ denotes random normal noise simulated with mean 0 and standard deviation 0.5. We take a pilot sample consisting of 4,000 control subjects, and an analysis sample with 3,000 subjects, among whom 292 receive treatment.

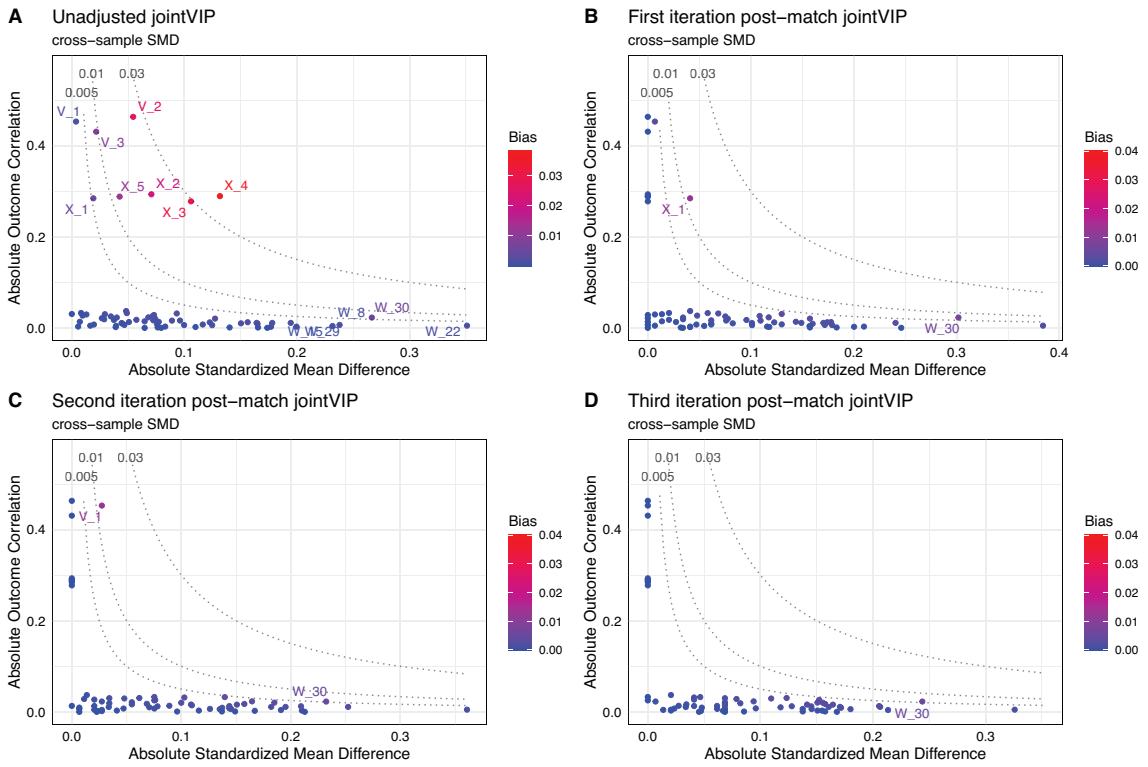
A.3.2 Design and estimation

In the simulated dataset, we conduct optimal pair matching using a Mahalanobis distance[51]. We conduct a randomization test for the difference-in-mean statistics using a similar formal framework to the one described in Section 3.3 of the main manuscript, but also invert this test to construct confidence intervals (CIs) for matched pairs, following Rosenbaum[93]).

The key question in this study design is which variables to use when computing the multivariate matching distance. We test three general strategies: using all available variables, selecting variables based on imbalance information alone, and selecting variables using jointVIP. For the latter two approaches, we also consider successive refinements of an initial match based on computation of post-match versions of the relevant diagnostic.

For adjustment based on imbalance alone, we first generate a traditional Love plot or balance table using pooled standardized mean difference (SMD) to evaluate imbalance. There are 68 variables and 24 variables are have measured imbalance above the traditional absolute 0.1 cutoff for pooled SMD. First adjustment for the 24 variables would still leave 15 variables still imbalanced, including 12 not included in the first Mahalanobis distance. Refining the original distance to include these 12 additional variables leaves 18 variables still imbalanced, but all are already present in the Mahalanobis distance so no further refinements are explored.

For adjustment using jointVIP, the researcher would first examine the unadjusted jointVIP (Supplemental Fig. A.2.A). The unadjusted plot indicates 6 variables needing adjustment above a 0.01 bias tolerance threshold. After the initial adjustment, the post-match jointVIP



Appendix Figure A.2: Iterative usage of the joint variable importance plot showing all the variables are under 0.005 bias curve

(Supplemental Fig. A.2.B) indicates an additional variable to be included as a variable for tuning using this bias threshold, and a third iteration is suggested by the post-match jointVIP following this refinement (Supplemental Fig. A.2.C). Note the difference in approach between jointVIP-based and imbalance-based selection; although unadjusted bias metrics are under 0.005 for every variable after the final jointVIP refinement (Supplemental Fig. A.2.D), 28 absolute standardized mean differences remain above the 0.1 cutoff.

Point estimates and confidence intervals for all six matches are reported in Table A.1. While imbalance-based selection improves on the strategy using all variables (for which the confidence interval does not even cover the true parameter), jointVIP is by far the best performer both in terms of smallest bias achieved and shortest confidence interval constructed. Replication code for the simulation is publicly available on GitHub: (https://github.com/ldliao/jointVIP/blob/main/paper/simulation/code).

Matched Pairs Design	**Estimate and CI**
Adjust all background variables	0.086 CI:(-0.327, 0.499)

Adjust with imbalance via pooled SMD	first iteration	0.349 CI:(-0.177, 0.862)
	second iteration	0.155 CI:(-0.188, 0.886)
Adjust with bias via jointVIP	first iteration	0.555 CI:(0.239, 0.870)
	second iteration	0.515 CI:(0.259, 0.770)
	third iteration	**0.532** CI:(0.4545, 0.6099)

Appendix Table A.1: Comparing different designs in the simulation, where the true treatment effect is 0.5.

A.4 Summary of all baseline variables of pregnant individuals with gestational diabetes

Appendix Table A.2: Summary of all baseline variables of pregnant individuals with gestational diabetes.

		2007-2010 Control n = 7,526	2011-2021 Control n = 19,183	2011-2021 Treated n = 10,786
Age (%)	Under 25	552 (7.3)	1,029 (5.4)	349 (3.2)
	Between 25-29	1,665 (22.1)	3,762 (19.6)	1,866 (17.3)
	Between 30-34	2,584 (34.3)	7,101 (37.0)	4,165 (38.6)
	Over 35	2,725 (36.2)	7,291 (38.0)	4,406 (40.8)
KP member 6 months prior to pregnancy = yes (%)		5,891 (78.3)	15,752 (82.1)	9,004 (83.5)
Median housing income (%)	Less than $ 40,000	892 (11.9)	784 (4.1)	355 (3.3)
	$ 40,000 - $ 59,999	1,800 (23.9)	2,807 (14.6)	1,325 (12.3)
	$ 60,000 - $ 79,999	1,903 (25.3)	3,769 (19.6)	2,132 (19.8)
	$ 80,000 and above	2,931 (38.9)	11,823 (61.6)	6,974 (64.7)
Parity (%)	0	3,121 (41.5)	7,611 (39.7)	3,873 (35.9)
	1	2,347 (31.2)	6,566 (34.2)	3,994 (37.0)
	2 or more	2,058 (27.3)	5,006 (26.1)	2,919 (27.1)
Pre-pregnancy BMI (%)	Underweight	100 (1.3)	391 (2.0)	76 (0.7)
	Normal	1,921 (25.5)	5,107 (26.6)	1,706 (15.8)
	Overweight	2,847 (37.8)	6,369 (33.2)	3,361 (31.2)
	Obese	2,658 (35.3)	7,316 (38.1)	5,643 (52.3)
Race/ethnicity (%)	Asian or Pacific Islander	2,919 (38.8)	8,553 (44.6)	4,560 (42.3)
	Hispanic (%)	2,322 (30.9)	5,013 (26.1)	2,923 (27.1)
	White (%)	1,602 (21.3)	4,090 (21.3)	2,381 (22.1)
	Black or African American	315 (4.2)	736 (3.8)	410 (3.8)
	Other/Unknown (%)	368 (4.9)	791 (4.1)	512 (4.7)
Singleton pregnancy = yes (%)		7,286 (96.8)	18,579 (96.9)	10,567 (98.0)

Continuation of Table A.2

	2007-2010 Control n = 7,526	2011-2021 Control n = 19,183	2011-2021 Treated n = 10,786
Alcohol consumption prior to pregnancy = yes (%)	2,915 (38.7)	7,715 (40.2)	4,954 (45.9)
Alcohol consumption during pregnancy = yes (%)	437 (5.8)	1,893 (9.9)	946 (8.8)
Smoking prior to pregnancy = yes (%)	611 (8.1)	919 (4.8)	690 (6.4)
Smoking during pregnancy = yes (%)	207 (2.8)	314 (1.6)	222 (2.1)
Arrythmia diagnosis = yes (%)	52 (0.7)	194 (1.0)	107 (1.0)
Asthma diagnosis = yes (%)	737 (9.8)	2,279 (11.9)	1,507 (14.0)
Chronic hypertension = yes (%)	435 (5.8)	846 (4.4)	591 (5.5)
Depression diagnosis prior to pregnancy = yes (%)	624 (8.3)	1,538 (8.0)	1,020 (9.5)
Depression diagnosis during pregnancy = yes (%)	455 (6.0)	2,639 (13.8)	1,626 (15.1)
Dyslipidemia diagnosis = yes (%)	305 (4.1)	1173 (6.1)	861 (8.0)
Family history of diabetes = yes (%)	198 (2.6)	1202 (6.3)	822 (7.6)
History of abortive outcome = yes (%)	694 (9.2)	2,253 (11.7)	1,412 (13.1)
History of GDM = yes (%)	856 (11.4)	3,401 (17.7)	2,608 (24.2)
History of macrosomia = yes (%)	69 (0.9)	145 (0.8)	147 (1.4)
Polycystic ovary syndrome by diagnosis = yes (%)	263 (3.5)	831 (4.3)	653 (6.1)
Pre-pregnancy pre-diabetes = yes (%)	479 (6.4)	1,802 (9.4)	1,915 (17.8)
Count of blood pressure measurements taken (mean (SD))	6.21 (9.35)	4.65 (3.49)	5.11 (3.83)
Diastolic blood pressure prior to pregnancy (mean (SD))	72.89 (8.32)	71.44 (10.15)	72.47 (10.08)
Systolic blood pressure prior to pregnancy (mean (SD))	116.69 (11.53)	117.26 (13.38)	118.86 (13.36)
Diastolic blood pressure prior to GDM diagnosis (mean (SD))	68.63 (8.76)	66.82 (9.70)	67.83 (9.59)
Systolic blood pressure prior to GDM diagnosis (mean (SD))	114.45 (12.17)	114.43 (13.14)	116.37 (12.94)
Average diastolic blood pressure prior to pregnancy (mean (SD))	72.96 (7.38)	71.26 (8.05)	72.24 (8.06)
Average systolic blood pressure prior to pregnancy (mean (SD))	117.33 (10.42)	117.14 (11.00)	118.70 (10.98)
Median diastolic blood pressure prior to pregnancy (mean (SD))	72.98 (7.52)	71.22 (8.26)	72.21 (8.29)
Median systolic blood pressure prior to pregnancy (mean (SD))	117.16 (10.56)	116.86 (11.23)	118.41 (11.25)
Average diastolic blood pressure prior to GDM (mean (SD))	69.26 (7.37)	67.68 (8.08)	68.89 (8.18)
Average systolic blood pressure prior to GDM (mean (SD))	115.11 (10.55)	115.38 (11.15)	117.40 (11.17)

Continuation of Table A.2

		2007-2010 Control n = 7,526	2011-2021 Control n = 19,183	2011-2021 Treated n = 10,786
Median diastolic blood pressure prior to GDM (mean (SD))		69.20 (7.55)	67.56 (8.29)	68.80 (8.38)
Median systolic blood pressure prior to GDM (mean (SD))		114.91 (10.73)	115.13 (11.38)	117.16 (11.40)
Infant sex (%)	Female	3,577 (47.5)	9,250 (48.2)	5,145 (47.7)
	Male	3,855 (51.2)	9,862 (51.4)	5,605 (52.0)
	Unknown	94 (1.2)	71 (0.4)	36 (0.3)
Glucose challenge test value (mean (SD))		169.43 (22.38)	169.71 (22.14)	173.22 (24.32)
Gestational age at GDM diagnosis (mean (SD))		26.06 (6.10)	26.86 (6.02)	23.53 (7.12)
Gestational weight gain up to GDM diagnosis (mean (SD))		15.28 (11.04)	14.82 (10.99)	12.86 (11.35)
Gestational hypertension = yes (%)		385 (5.1)	1,546 (8.1)	884 (8.2)
GDM diagnosed by the C-C criteria = yes (%)		7,323 (97.3)	17,303 (90.2)	8,419 (78.1)
One-hour OGTT = abnormal (%)		6,536 (86.8)	15,534 (81.0)	8,244 (76.4)
Two-hour OGTT = abnormal (%)		6,662 (88.5)	15,829 (82.5)	7,082 (65.7)
Three-hour OGTT = abnormal (%)		2,875 (38.2)	7,194 (37.5)	2,966 (27.5)
OGTT for fasting blood glucose = abnormal (%)		2,165 (28.8)	3,762 (19.6)	4,512 (41.8)
GDM severity = severe (%)		775 (10.3)	1,744 (9.1)	1,215 (11.3)

BMI: body mass index, C-C: Carpenter-Coustan, GDM: gestational diabetes, KP: Kaiser Permanente, OGTT: oral glucose tolerance test, SD: standard deviation.
A normal OGTT fasting blood glucose level is lower than 95 mg/dL.
A normal one-hour OGTT blood glucose level is lower than 180 mg/dL.
A normal two-hour OGTT blood glucose level is lower than 155 mg/dL.
A normal three-hour OGTT blood glucose level is lower than 140 mg/dL.

A.5 Missingness summary and out-of-bag imputation error estimate

Appendix Table A.3: Missingness summary and out-of-bag imputation error estimate.

Variables with missingness indicators

	2007-2010 Control n = 7,526	2011-2021 Control n = 19,183	2011-2021 Treated n = 10,786
Gestational weight gain up to GDM diagnosis (%)	1,391 (18.5)	2,504 (13.1)	1,307 (12.1)
Blood pressure measured prior to pregnancy (%)	2,988 (39.7)	2,914 (15.2)	1,353 (12.5)
Blood pressure measured prior to GDM (%)	658 (8.7)	35 (0.2)	46 (0.4)
Pre-pregnancy BMI (%)	1,326 (17.6)	866 (4.5)	350 (3.2)
Glucose challenge test value (%)	204 (2.7)	152 (0.8)	104 (1.0)
Median housing income	4 (0.1)	7 (0.0)	0 (0.0)
Parity[1]	0 (0.0)	26 (0.0)	13 (0.0)

Out-of-bag imputation error

Year	NRMSE	PFC
2007	$2.14*10^{-6}$	$7.96*10^{-2}$
2008	$2.27*10^{-6}$	0.00
2009	$2.35*10^{-6}$	$8.02*10^{-2}$
2010	$2.35*10^{-6}$	$7.83*10^{-2}$
2011	$3.83*10^{-5}$	$1.01*10^{-1}$
2012	$4.17*10^{-5}$	0.00
2013	$4.09*10^{-5}$	$9.73*10^{-2}$
2014	$4.08*10^{-5}$	0.00
2015	$4.13*10^{-5}$	0.00
2016	$4.13*10^{-5}$	$1.69*10^{-1}$
2017	$4.04*10^{-5}$	$8.37*10^{-2}$
2018	$3.98*10^{-5}$	$1.76*10^{-1}$
2019	$3.79*10^{-5}$	0.00
2020	$3.43*10^{-5}$	$1.82*10^{-1}$
2021	$3.28*10^{-5}$	$1.82*10^{-1}$

NOTE: BMI: body mass index, GDM: gestational diabetes, NRMSE: Root mean squared error, PFC: proportion of falsely classified entries

[1]Parity is only missing in analysis 2011-2021 dataset. Since the missingness is quite small compared to data available, this indicator is dropped after imputation.

2007-2010 is the pilot data, and 2011-2021 is the analysis data.

The out-of-bag error imputation is calculated separately for continuous (NRMSE) and categorical variables (PFC).

A.6 Pre-and-post match comparison of background variables with high unadjusted bias

Appendix Table A.4: Pre-and-post match comparison of background variables with high unadjusted bias.

Caesarean section delivery

Background variable	Pre-match bias	Post-match bias
Propensity score[1]	0.074	0.0056
Prognostic score	0.043	0.0025
OGTT for fasting blood glucose	0.023	0.0000
Obese pre-pregnancy BMI	0.023	0.0000
GDM diagnosed by the C-C criteria	0.017	0.0001
Gestational age at GDM diagnosis	0.013	0.0018
Pre-pregnancy pre-diabetes	0.012	0.0005
Normal pre-pregnancy BMI	0.010	0.0001

NOTE: BMI: body mass index, C-C: Carpenter-Coustan, C-section: Cesarean section, GDM: gestational diabetes, OGTT: oral glucose tolerance test
[1]Denotes the maximum post-match unadjusted bias for that outcome.

A.7 Summary of all baseline variables of post-match treated pregnant individuals

Appendix Table A.5: Summary of all baseline variables of post-match treated pregnant individuals.

		C-section delivery included n = 8,693	C-section delivery excluded n = 2,093
Age (%)	Under 25	322 (3.7)	27 (1.3)
	Between 25-29	1,522 (17.5)	344 (16.4)
	Between 30-34	3,405 (39.2)	760 (36.3)
	Over 35	3,444 (39.6)	962 (46.0)
KP member 6 months prior to pregnancy = yes (%)		7,229 (83.2)	1,775 (84.8)
Median housing income (%)	Less than $ 40,000	286 (3.3)	69 (3.3)
	$ 40,000 - $ 59,999	1,051 (12.1)	274 (13.1)
	$ 60,000 - $ 79,999	1,720 (19.8)	412 (19.7)

Continuation of Table A.5

		C-section delivery included n = 8,693	C-section delivery excluded n = 2,093
	$ 80,000 and above	5,636 (64.8)	1,338 (63.9)
Parity (%)	0	3,214 (37.0)	659 (31.5)
	1	3,203 (36.8)	791 (37.8)
	2 or more	2,276 (26.2)	643 (30.7)
Pre-pregnancy BMI (%)	Underweight	69 (0.8)	7 (0.3)
	Normal	1,612 (18.5)	94 (4.5)
	Overweight	2,862 (32.9)	499 (23.8)
	Obese	4,150 (47.7)	1,493 (71.3)
Race/ethnicity (%)	Asian or Pacific Islander	3,796 (43.7)	764 (36.5)
	Hispanic (%)	2,316 (26.6)	607 (29.0)
	White (%)	1,880 (21.6))	501 (23.9)
	Black or African American	315 (3.6)	95 (4.5)
	Other/Unknown (%)	386 (4.4)	126 (6.0)
Singleton pregnancy = yes (%)		8,508 (97.9)	2,059 (98.4)
Alcohol consumption prior to pregnancy = yes (%)		3,731 (42.9)	1,223 (58.4)
Alcohol consumption during pregnancy = yes (%)		750 (8.6)	196 (9.4)
Smoking prior to pregnancy = yes (%)		467 (5.4)	223 (10.7)
Smoking during pregnancy = yes (%)		156 (1.8)	66 (3.2)
Arrythmia diagnosis = yes (%)		81 (0.9)	26 (1.2)
Asthma diagnosis = yes (%)		1,138 (13.1)	369 (17.6)
Chronic hypertension = yes (%)		397 (4.6)	194 (9.3)
Depression diagnosis prior to pregnancy = yes (%)		786 (9.0)	234 (11.2)
Depression diagnosis during pregnancy = yes (%)		1,230 (14.1)	396 (18.9)
Dyslipidemia diagnosis = yes (%)		619 (7.1)	242 (11.6)
Family history of diabetes = yes (%)		590 (6.8)	232 (11.1)
History of abortive outcome = yes (%)		1,102 (12.7)	310 (14.8)
History of GDM = yes (%)		1,930 (22.2)	678 (32.4)
History of macrosomia = yes (%)		105 (1.2)	42 (2.0)
Polycystic ovary syndrome by diagnosis = yes (%)		468 (5.4)	185 (8.8)
Pre-pregnancy pre-diabetes = yes (%)		1,213 (14.0)	702 (33.5)
Count of blood pressure measurements taken (mean (SD))		4.95 (3.58)	5.77 (4.67)

Continuation of Table A.5

		C-section delivery included n = 8,693	C-section delivery excluded n = 2,093
Diastolic blood pressure prior to pregnancy (mean (SD))		72.05 (9.99)	74.21 (10.27)
Systolic blood pressure prior to pregnancy (mean (SD))		118.10 (13.26)	121.97 (13.30)
Diastolic blood pressure prior to GDM diagnosis (mean (SD))		67.29 (9.40)	70.04 (10.04)
Systolic blood pressure prior to GDM diagnosis (mean (SD))		115.50 (12.67)	119.96 (13.46)
Average diastolic blood pressure prior to pregnancy (mean (SD))		71.77 (7.93)	74.15 (8.29)
Average systolic blood pressure prior to pregnancy (mean (SD))		117.91 (10.81)	121.98 (11.07)
Median diastolic blood pressure prior to pregnancy (mean (SD))		71.75 (8.16)	74.15 (8.54))
Median systolic blood pressure prior to pregnancy (mean (SD))		117.62 (11.06)	121.71 (11.42)
Average diastolic blood pressure prior to GDM (mean (SD))		68.36 (7.96)	71.06 (8.71)
Average systolic blood pressure prior to GDM (mean (SD))		116.54 (10.92)	120.99 (11.49)
Median diastolic blood pressure prior to GDM (mean (SD))		68.27 (8.16)	70.98 (8.89)
Median systolic blood pressure prior to GDM (mean (SD))		116.27 (11.15)	120.85 (11.67)
Infant sex (%)	Female	4,134 (47.6)	1,011 (48.3)
	Male	4528 (52.1)	1077 (51.5)
	Unknown	31 (0.4)	5 (0.2)
Glucose challenge test value (mean (SD))		172.47 (23.80)	176.32 (26.16)
Gestational age at GDM diagnosis (mean (SD))		24.52 (6.57)	19.39 (7.80)
Gestational weight gain up to GDM diagnosis (mean (SD))		13.63 (11.16)	9.64 (11.56)
Gestational hypertension = yes (%)		634 (7.3)	250 (11.9)
GDM diagnosed by the C-C criteria = yes (%)		7,107 (81.8)	1,312 (62.7)
One-hour OGTT = abnormal (%)		6,825 (78.5)	1,419 (67.8)
Two-hour OGTT = abnormal (%)		6,146 (70.7)	936 (44.7)
Three-hour OGTT = abnormal (%)		2,575 (29.6)	391 (18.7)
OGTT for fasting blood glucose = abnormal (%)		3,095 (35.6)	1,417 (67.7)
GDM severity = severe (%)		936 (10.8)	279 (13.3)

Appendix B Supplementary Material for Chapter 3

B.1 Complete table of baseline variables and preexisting conditions

Appendix Table B.1: Complete table of baseline variables and preexisting conditions.

	All time (2020/03-2021/11)	Phase 1 (2020/03-2020/10)	Phase 2 (2020/11-2021/03)	Phase 3 (2021/04-2021/11)
sample size	1,423,720	303,278	425,698	694,744
Demographic variables				
Age in years (mean (SD))	42.15 (15.70)	46.41 (16.04)	44.89 (16.27)	38.61 (14.34)
Sex = male (%)	729,782 (51.3)	158,248 (52.2)	218,165 (51.2)	353,369 (50.9)
Insured by IMSS = 1 (%))	1,358,440 (95.4)	288,588 (95.2)	402,754 (94.6)	667,098 (96.0)
Indigenous = 1 (%)	7,381 (0.5)	2,200 (0.7)	1,628 (0.4)	3,553 (0.5)
Year-month patient initiated care				
2020/03	1,061 (0.1)	1,061 (0.3)	0 (0.0)	0 (0.0)
2020/04	10,832 (0.8)	10,832 (3.6)	0 (0.0)	0 (0.0)
2020/05	30,720 (2.2)	30,720 (10.1)	0 (0.0)	0 (0.0)
2020/06	51,079 (3.6)	51,079 (16.8)	0 (0.0)	0 (0.0)
2020/07	60,780 (4.3)	60,780 (20.0)	0 (0.0)	0 (0.0)
2020/08	49,618 (3.5)	49,618 (16.4)	0 (0.0)	0 (0.0)
2020/09	44,758 (3.1)	44,758 (14.8)	0 (0.0)	0 (0.0)
2020/10	54,430 (3.8)	54,430 (17.9)	0 (0.0)	0 (0.0)
2020/11	65,437 (4.6)	0 (0.0)	65,437 (15.4)	0 (0.0)
2020/12	93,748 (6.6)	0 (0.0)	93,748 (22.0)	0 (0.0)
2021/01	145,858 (10.2)	0 (0.0)	145,858 (34.3)	0 (0.0)
2021/02	68,421 (4.8)	0 (0.0)	68,421 (16.1)	0 (0.0)
2021/03	52,234 (3.7)	0 (0.0)	52,234 (12.3)	0 (0.0)
2021/04	35,181 (2.5)	0 (0.0)	0 (0.0)	35,181 (5.1)
2021/05	26,300 (1.8)	0 (0.0)	0 (0.0)	26,300 (3.8)
2021/06	45,986 (3.2)	0 (0.0)	0 (0.0)	45,986 (6.6)
2021/07	170,212 (12.0)	0 (0.0)	0 (0.0)	170,212 (24.5)
2021/08	249,477 (17.5)	0 (0.0)	0 (0.0)	249,477 (35.9)
2021/09	116,569 (8.2)	0 (0.0)	0 (0.0)	116,569 (16.8)
2021/10	48,515 (3.4)	0 (0.0)	0 (0.0)	48,515 (7.0)
2021/11	2,504 (0.2)	0 (0.0)	0 (0.0)	2,504 (0.4)
Mexican states (%)				
Aguascalientes	26,420 (1.9)	6,897 (2.3)	12,350 (2.9)	7,173 (1.0)
Baja California	43,925 (3.1)	13,677 (4.5)	14,188 (3.3)	16,060 (2.3)
Baja California Sur	24,521 (1.7)	4,300 (1.4)	5,423 (1.3)	14,798 (2.1)
Campeche	9,557 (0.7)	1,728 (0.6)	765 (0.2)	7,064 (1.0)
CDMX 1 Noroeste	32,552 (2.3)	5,374 (1.8)	13,174 (3.1)	14,004 (2.0)
CDMX 2 Noreste	54,249 (3.8)	11,370 (3.7)	20,273 (4.8)	22,606 (3.3)
CDMX 3 Suroeste	42,896 (3.0)	9,701 (3.2)	17,588 (4.1)	15,607 (2.2)

Continuation of Table B.1

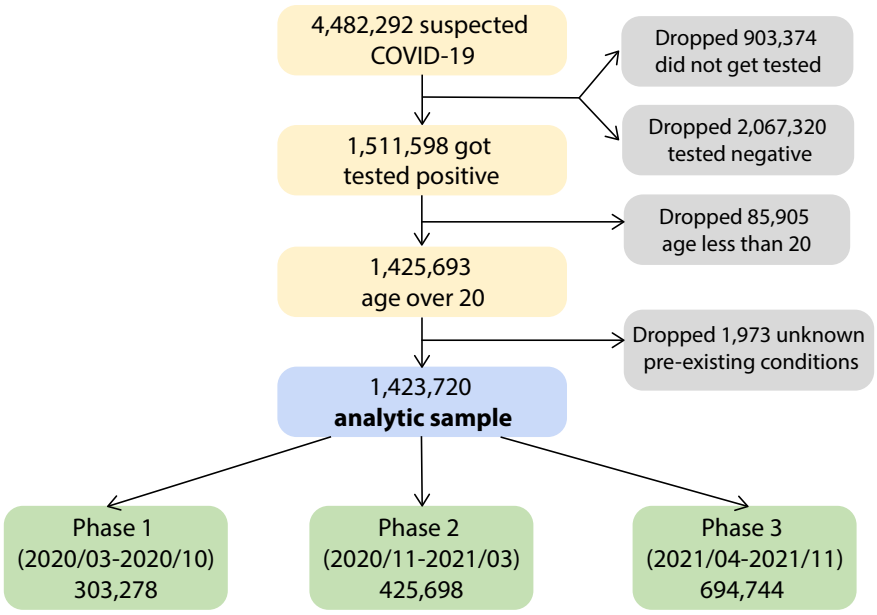
	All time (2020/03-2021/11)	Phase 1 (2020/03-2020/10)	Phase 2 (2020/11-2021/03)	Phase 3 (2021/04-2021/11)
CDMX 4 Sureste	62,097 (4.4)	13,248 (4.4)	25,899 (6.1)	22,950 (3.3)
Chiapas	14,826 (1.0)	2,836 (0.9)	1,801 (0.4)	10,189 (1.5)
Chihuahua	23,229 (1.6)	6,489 (2.1)	6,879 (1.6)	9,861 (1.4)
Coahuila	48,933 (3.4)	16,355 (5.4)	15,459 (3.6)	17,119 (2.5)
Colima	18,310 (1.3)	2,997 (1.0)	2,633 (0.6)	12,680 (1.8)
Durango	19,738 (1.4)	6,228 (2.1)	6,674 (1.6)	6,836 (1.0)
Guanajuato	61,570 (4.3)	11,595 (3.8)	29,274 (6.9)	20,701 (3.0)
Guerrero	23,871 (1.7)	4,502 (1.5)	3,920 (0.9)	15,449 (2.2)
Hidalgo	23,673 (1.7)	4,658 (1.5)	8,266 (1.9)	10,749 (1.5)
Jalisco	104,054 (7.3)	19,491 (6.4)	27,868 (6.5)	56,695 (8.2)
Mexico Oriente	108,067 (7.6)	20,368 (6.7)	39,633 (9.3)	48,066 (6.9)
Mexico Poniente	48,973 (3.4)	11,632 (3.8)	17,026 (4.0)	20,315 (2.9)
Michoacan	30,570 (2.1)	6,246 (2.1)	7,221 (1.7)	17,103 (2.5)
Morelos	18,797 (1.3)	2,845 (0.9)	6,817 (1.6)	9,135 (1.3)
Nayarit	23,934 (1.7)	3,378 (1.1)	2,994 (0.7)	17,562 (2.5)
Nuevo Leon	105,912 (7.4)	23,776 (7.8)	30,114 (7.1)	52,022 (7.5)
Oaxaca	24,324 (1.7)	4,493 (1.5)	5,391 (1.3)	14,440 (2.1)
Puebla	50,998 (3.6)	9,287 (3.1)	17,932 (4.2)	23,779 (3.4)
Queretaro	41,977 (2.9)	4,707 (1.6)	19,259 (4.5)	18,011 (2.6)
Quintana Roo	38,390 (2.7)	4,607 (1.5)	4,542 (1.1)	29,241 (4.2)
San Luis Potosi	26,118 (1.8)	7,353 (2.4)	7,573 (1.8)	11,192 (1.6)
Sinaloa	44,333 (3.1)	11,030 (3.6)	9,405 (2.2)	23,898 (3.4)
Sonora	27,691 (1.9)	7,245 (2.4)	6,083 (1.4)	14,363 (2.1)
Tabasco	16,004 (1.1)	2,622 (0.9)	1,719 (0.4)	11,663 (1.7)
Tamaulipas	38,941 (2.7)	8,504 (2.8)	7,442 (1.7)	22,995 (3.3)
Tlaxcala	13,809 (1.0)	2,769 (0.9)	4,957 (1.2)	6,083 (0.9)
Veracruz Norte	41,804 (2.9)	10,002 (3.3)	6,801 (1.6)	25,001 (3.6)
Veracruz Sur	35,019 (2.5)	10,292 (3.4)	4,970 (1.2)	19,757 (2.8)
Yucatan	35,126 (2.5)	5,968 (2.0)	5,091 (1.2)	24,067 (3.5)
Zacatecas	18,512 (1.3)	4,708 (1.6)	8,294 (1.9)	5,510 (0.8)
Preexisting conditions				
Asthma = yes (%)	25,297 (1.8)	7,951 (2.6)	7,765 (1.8)	9,581 (1.4)
Cardiovascular patient disease = yes (%)	17,816 (1.3)	6,643 (2.2)	6,389 (1.5)	4,784 (0.7)
Chronic liver disease = yes (%)	1,875 (0.1)	710 (0.2)	668 (0.2)	497 (0.1)
COPD = yes (%)	15,390 (1.1)	5,825 (1.9)	5,496 (1.3)	4,069 (0.6)
Diabetes = yes (%)	169,869 (11.9)	55,551 (18.3)	61,120 (14.4)	53,198 (7.7)
Hemolytic anemia = yes (%)	705 (0.0)	276 (0.1)	246 (0.1)	183 (0.0)
HIV = yes (%)	4,717 (0.3)	1,133 (0.4)	1,425 (0.3)	2,159 (0.3)
Hypertension = yes (%)	228,901 (16.1)	72,615 (23.9)	83,735 (19.7)	72,551 (10.4)
Immunosuppression = yes (%)	10,434 (0.7)	4,102 (1.4)	3,453 (0.8)	2,879 (0.4)
Neurological disease = yes (%)	1,645 (0.1)	544 (0.2)	559 (0.1)	542 (0.1)
Obesity = yes (%)	181,736 (12.8)	55,965 (18.5)	60,217 (14.1)	65,554 (9.4)
Smoking = yes (%)	87,161 (6.1)	21,253 (7.0)	28,346 (6.7)	37,562 (5.4)

Continuation of Table B.1

	All time (2020/03-2021/11)	Phase 1 (2020/03-2020/10)	Phase 2 (2020/11-2021/03)	Phase 3 (2021/04-2021/11)
Cancer diagnosis = yes (%)	3,751 (0.3)	1,178 (0.4)	1,317 (0.3)	1,256 (0.2)
Renal disease diagnosis = yes (%)	24,099 (1.7)	8,912 (2.9)	8,555 (2.0)	6,632 (1.0)
Tuberculosis = yes (%)	675 (0.0)	203 (0.1)	218 (0.1)	254 (0.0)

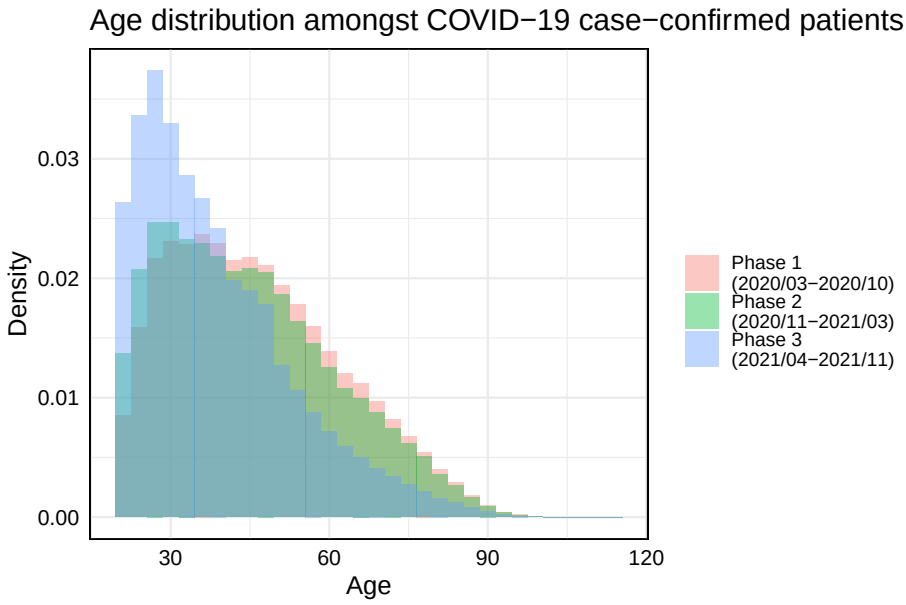
CDMX: Ciudad de México; COPD: Chronic obstructive pulmonary disease;
HIV: human immunodeficiency virus; SD: standard deviation.
Mexican states refer to where the patient was treated.

B.2 Flowchart for analytic sample development



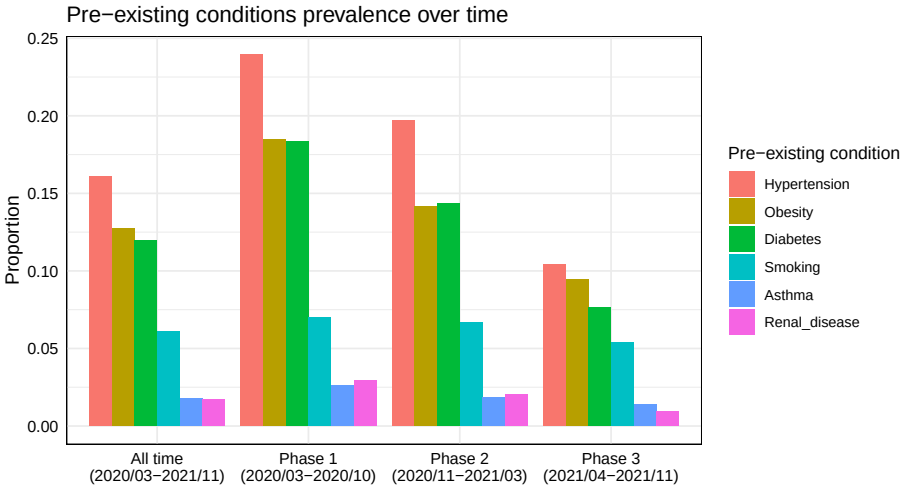
Appendix Figure B.1: Flowchart for analytic sample development.

B.3 Age distribution for laboratory-confirmed COVID-19 patients



Appendix Figure B.2: Age distribution for laboratory-confirmed COVID-19 patients.

B.4 Prevalence of preexisting conditions prevalence over time



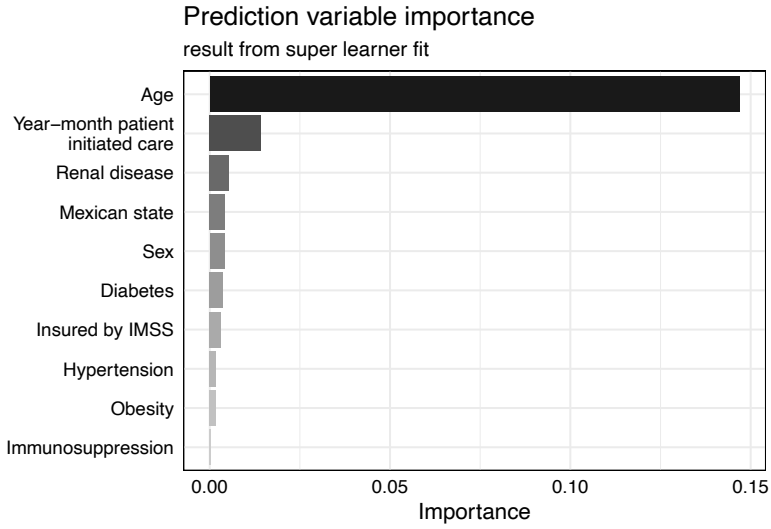
Appendix Figure B.3: Prevalence of preexisting conditions prevalence over time.

B.5 Weighted combination of the super learner fit

Machine learning candidate algorithm	Weights	Mean squared error	Standard error
Bayesian additive regression trees	0	0.266	0.0002
Bayesian generalized linear model	0	0.067	0.0003
Elastic net regression	0	0.068	0.0004
Empirical mean	0	0.094	0.0005
XGBoost (multiple tuning)	0.596 (combined)	0.065 (on average)	0.0003 (on average)
Generalized additive model	0.222	0.066	0.0004
LASSO regression	0	0.067	0.0004
Logistic regression	0	0.067	0.0004
Multivariate Adaptive Regression Splines	0	0.068	0.0004
Random forest	0.181	0.066	0.0002
Ridge regression	0	0.067	0.0002

Appendix Table B.2: Weighted combination of the super learner fit

B.6 Prediction variable importance predicted using the super learner fit



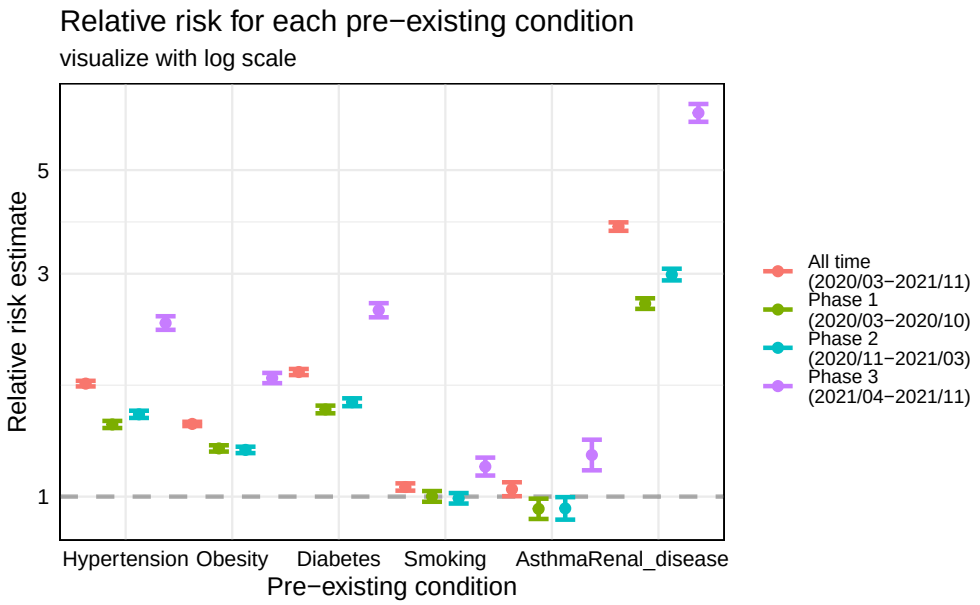
Appendix Figure B.4: Prediction variable importance predicted using the super learner fit.

B.7 Top 5 ranked most important variables for prediction

	All time (2020/03-2021/11)	Phase 1 (2020/03-2020/10)	Phase 2 (2020/11-2021/03)	Phase 3 (2021/04-2021/11)
Rank 1	Age 0.147	Age 0.209	Age 0.208	Age 0.069
Rank 2	Year-month patient initiated care 0.014)	Renal disease 0.008	Mexican state 0.007	Renal disease 0.004
Rank 3	Renal disease 0.005	Sex 0.007	Renal disease 0.006	Mexican state 0.003
Rank 4	Mexican state 0.004	Year-month patient initiated care 0.007)	Insured by IMSS 0.006	Diabetes 0.003
Rank 5	Sex 0.004	Mexican state 0.006	Sex 0.006	Insured by IMSS 0.002

Appendix Table B.3: Prediction results.

B.8 Relative risk for each preexisting condition associated with mortality



Appendix Figure B.5: Relative risk for each preexisting condition associated with mortality.

B.9 Targeted maximum likelihood estimation adjusted mortality risk, with or without the preexisting condition

	All time (2020/03-2021/11)		Phase 1 (2020/03-2020/10)		Phase 2 (2020/11-2021/03)		Phase 3 (2021/04-2021/11)	
	with	without	with	without	with	without	with	without
Renal disease	0.381	0.101	0.439	0.170	0.425	0.142	0.305	0.046
Diabetes	0.173	0.094	0.247	0.161	0.214	0.135	0.104	0.041
Hypertension	0.162	0.093	0.231	0.162	0.201	0.134	0.097	0.041
Obesity	0.141	0.099	0.212	0.168	0.177	0.141	0.080	0.045
Smoking	0.110	0.105	0.176	0.176	0.146	0.147	0.056	0.048
Asthma	0.109	0.105	0.166	0.177	0.139	0.147	0.059	0.049

Appendix Table B.4: Targeted maximum likelihood estimation adjusted mortality risk, with or without the pre-existing condition.

Appendix C Supplementary Material for Chapter 4

C.1 Expectation calculation when incorporating unobserved covariate

By definition, an unobserved covariate U is never seen in real data but we include such a variable in simulation. We aim to demonstrate that even if the outcome model can be learned perfectly from historical data, if there exists an unobserved shift between the historical and trial sample, then the learned historical outcome model (prognostic model) can never be equivalent to the trial outcome model. We explicitly write out the expectation of the outcome Y given treatment A, baseline covariates W, and data set indicator D using an unobserved covariate U.

$$E[Y|A,W,D=d] = \int E[Y|A,W,U=u,D=d]p(u|A,W,D=d)\mathrm{d}u$$

A shift in the distribution $p(u|A,W,D=1) \neq p(u|A,W,D=0)$ of the unobserved covariate U will generally result in unequal conditional expectations, i.e., $E[Y|A,W,D=1] \neq E[Y|A,W,D=0]$.

This shift is the basis of the simulation for Figure 3.B. In our simulation, we have $U|A,W,D=1 \sim \text{Unif}(0,1)$ and $U|A,W,D=0 \sim \text{Unif}(\underline{u},\overline{u})$ where the limits $underline(u), overline(u)$ increase or decrease past 0 or 1 depending on the desired magnitude of covariate shift. The "oracle" prognostic score is given by $E[Y|A,W,D=1]$, i.e. always integrating over the correct (trial) density $U|A,W,D=1 \sim \text{Unif}(0,1)$.

By framing shifts in the conditional mean as shifts in an unobserved covariate we can directly control the magnitude of the change instead of manually specifying different conditional mean functions.

C.2 Discrete learner specifications for simulation and case study.

C.2.1 Discrete super learner specifications

Machine learning is performed through discrete super learner that the targeted maximum likelihood estimator internally leverages. For simplicity, the prognostic model is built using the discrete super learner as well. A discrete super learner selects from a set of candidate models (i.e., the library) to obtain a single, best prediction model via cross-validation. In this section, we describe the exact tuning parameters and set up for the simulation and case study.

C.2.2 Simulation set up

Cross-validation: 5-fold cross-validation is used to select the best candidate learner in the library for historical sample size 1,000, and 10-fold cross-validation for historical sample size less than 1,000.
Cross-fit: 5-fold Cross-fitting is employed.
Discrete super learner library: Multivariate Adaptive Regression Splines with the highest interaction to be to the 3rd degree, linear regression, extreme gradient boosting with specifications: learning rate 0.1, tree depth 3, crossed with trees specified 25 to 500 by 25 increments. Cases with *fitted* prognostic score include an augmented library that includes candidate learners with prognostic score in addition.
Discrete super learner specifications: loss function is specified to be the mean square error loss.

C.2.3 Case study set up

Cross-validation: 20-fold cross-validation is used to select the best candidate learner in the library.
Cross-fit: 20-fold Cross-fitting is employed.
Discrete super learner library: Multivariate Adaptive Regression Splines with the highest interaction to be to the 3rd degree, logistic regression, extreme gradient boosting with specifications: learning rate 0.1, tree depths 3, 5, and 10, crossed with trees specified 25 to 500 by 25 increments, random forest with trees specified 25 to 500 by 25 increments, k-nearest neighbor of specification 3, 4, 5, 7, and 9 number of nearest neighbors, k. Cases with *fitted*

prognostic score include an augmented library that includes candidate learners with prognostic score in addition to without.
Discrete super learner specifications: loss function is specified to be the mean log likelihood loss.
Selected prognostic model: The selected learner from 20-fold cross validation is a linear regression model for both the reanalyses with $n = 419$ and $n = 100$.

C.3 Simulation results for different data generation processes

Appendix Table C.1: Mean of empirically estimated bias, variance, and standard errors of them for the targeted maximum likelihood estimator with or without prognostic score across different DGPs. For all the scenarios the conditional means are shared with the heterogeneous effect DGP, except for the constant effect DGP. Unless otherwise specified $(\tilde{n}, n) = (1000, 250)$.

Scenario	Estimator *prog.*	Bias	Var.	SE bias	SE var.	RMSE	power	coverage
heterogeneous effect	TMLE *none*	0.009	5.691	0.180	0.055	2.380	0.645	0.960
	TMLE *fitted*	-0.064	4.482	0.158	0.005	2.113	0.720	0.975
	TMLE *oracle*	-0.069	4.341	0.152	0.003	2.080	0.745	0.985
	linear *none*	0.009	9.774	0.213	0.026	3.119	0.405	0.950
	linear *fitted*	-0.064	9.710	0.137	0.029	3.108	0.420	0.945
	linear *oracle*	-0.069	9.727	0.138	0.028	3.111	0.420	0.945
	unadjusted *none*	0.153	9.509	0.139	0.011	3.080	0.435	0.950
constant effect	TMLE *none*	0.027	0.119	-0.010	0.001	0.345	0.655	0.945
	TMLE *fitted*	0.025	0.067	0.028	0.000	0.260	0.790	0.970
	TMLE *oracle*	0.034	0.074	0.005	0.000	0.273	0.780	0.960
	linear *none*	0.032	0.427	0.014	0.001	0.653	0.240	0.940
	linear *fitted*	0.023	0.069	0.014	0.000	0.263	0.800	0.970
	linear *oracle*	0.025	0.060	0.020	0.000	0.246	0.840	0.970
	unadjusted *none*	0.067	0.857	-0.037	0.001	0.926	0.150	0.945
small observed shift	TMLE *none*	-0.001	5.699	0.178	0.055	2.381	0.640	0.960
	TMLE *fitted*	-0.096	4.851	0.187	0.038	2.199	0.715	0.970
	TMLE *oracle*	-0.074	4.411	0.133	0.003	2.096	0.740	0.985
	linear *none*	0.071	9.771	0.213	0.025	3.119	0.405	0.950
	linear *fitted*	0.094	9.835	0.205	0.026	3.130	0.405	0.950
	linear *oracle*	0.025	9.702	0.142	0.028	3.107	0.415	0.945
	unadjusted *none*	0.153	9.509	0.139	0.011	3.080	0.435	0.950
small unobserved shift	TMLE *none*	-0.032	5.803	0.155	0.050	2.403	0.640	0.950
	TMLE *fitted*	-0.026	5.346	0.108	0.015	2.306	0.695	0.975
	TMLE *oracle*	-0.076	4.434	0.129	0.003	2.102	0.745	0.985
	linear *none*	0.058	9.992	0.174	0.025	3.154	0.390	0.940
	linear *fitted*	-0.039	9.998	0.106	0.027	3.154	0.410	0.935
	linear *oracle*	-0.022	9.774	0.128	0.027	3.119	0.390	0.935
	unadjusted *none*	0.118	9.314	0.173	0.010	3.046	0.425	0.950
small historical sample $(\tilde{n}, n) = (100, 250)$	TMLE *none*	0.143	7.301	-0.154	0.039	2.699	0.690	0.930
	TMLE *fitted*	0.121	6.006	-0.074	0.017	2.448	0.735	0.925
	TMLE *oracle*	0.104	5.208	-0.046	0.004	2.279	0.755	0.935
	linear *none*	0.234	11.884	-0.097	0.027	3.447	0.440	0.935

Continuation of Table C.1

Scenario	Estimator *prog.*	Bias	Var.	SE bias	SE var.	RMSE	power	coverage
	linear *fitted*	0.090	11.499	-0.093	0.028	3.384	0.430	0.955
	linear *oracle*	0.093	11.309	-0.091	0.027	3.356	0.425	0.950
	unadjusted *none*	0.146	10.439	-0.009	0.010	3.226	0.425	0.960
small trial sample $(\tilde{n}, n) = (1000, 100)$	TMLE *none*	0.758	24.859	0.366	0.237	5.031	0.215	0.960
	TMLE *fitted*	0.155	14.698	-0.116	0.039	3.827	0.380	0.955
	TMLE *oracle*	0.092	14.016	-0.083	0.041	3.736	0.365	0.960
	linear *none*	0.647	30.222	0.208	0.285	5.522	0.200	0.940
	linear *fitted*	0.607	30.267	0.080	0.336	5.521	0.210	0.945
	linear *oracle*	0.620	30.279	0.088	0.326	5.524	0.205	0.950
	unadjusted *none*	0.670	21.590	0.468	0.076	4.683	0.205	0.970]

C.4 Case study data summary

Data name	**Trial ID**	**Duration**	**Titration target (mmol/L)**	**Blinding type**	**Number of participants**	
					randomized	completed
New RCT	NN9068-4229	26 weeks	4.0-5.0	Open-label	210	206
Historical	NN9068-4228	104 weeks	4.0-5.0	Open-label	504	481
	NN1250-3579	52 weeks	4.0-5.0	Open-label	257	197
	NN1250-3586	26 weeks	4.0-5.0	Open-label	146	136
	NN1250-3672	26 weeks	4.0-5.0	Open-label	230	201
	NN1250-3718	26 weeks	4.0-5.0	Open-label	234	209
	NN1250-3724	26 weeks	4.0-5.0	Open-label	230	206
	NN1250-3587	26 weeks	4.0-5.0	Open-label	278	254
	NN9535-3625	30 weeks	4.0-5.5	Open-label	365	343
	NN2211-1697	26 weeks	≤ 5.0	Double-blinded	34	219
	NN5401-3590	26 weeks	3.9-5.0	Open-label	264	232
	NN5401-3726	26 weeks	3.9-5.0	Open-label	extension of 3590	209
	NN5401-3896	26 weeks	3.9-5.0	Open-label	149	137
	NN1436-4383	26 weeks	4.4-7.2	Double-blinded	122	119
	NN1436-4465	16 weeks	4.4-7.2	Open-label	51	51
	NN1436-4477	78 weeks	4.4-7.2	Open-label	492	477

Appendix Table C.2: Summary of case study data provided by Novo Nordisk A/S. The new RCT data is highlighted in grey. The historical data consists of all the data sets that are not highlighted. The number of participants refers to the number of participants receiving the existing daily insulin treatment IGlar.

C.5 Summary of continuous measurements of the baseline

Appendix Table C.3: Summary of the continuous baseline covariates.

	Historical sample	New random trial sample
sample size	3311	419
age (years)		
N	3311	419
mean (SD)	57.34 (9.92)	56.67 (10.28)
median	58.00	58.00
min; max	21.00; 85.00	25.00; 83.00
alanine aminotransferase (U/L)		
N	3303	419
mean (SD)	29.51 (17.97)	26.63 (15.48)
median	25.00	23.00
min; max	2.50; 333.00	6.00; 138.00
albumin (g/dL)		
N	3306	419
mean (SD)	4.48 (0.28)	4.51 (0.25)
median	4.50	4.50
min; max	2.50; 5.90	3.80; 5.20
alkaline phosphatase (U/L)		
N	3305	419
mean (SD)	75.99 (23.50)	71.82 (22.31)
median	73.00	68.00
min; max	19.00; 261.00	20.00; 196.00
aspartate aminotransferase (U/L)		
N	3299	419
mean (SD)	23.22 (12.13)	21.38 (9.86)
median	20.00	19.00
min; max	6.00; 227.00	6.00; 89.00
basophils blood (%)		
N	3284	419
mean (SD)	0.53 (0.38)	0.39 (0.22)
median	0.40	0.40
min; max	0.00; 4.40	0.00; 1.60
body mass index (kg/m^2)		
N	3309	419
mean (SD)	30.72 (5.69)	31.22 (4.82)
median	30.22	31.00
min; max	16.01; 56.39	20.00; 43.30
body weight (kg)		
N	3309	419
mean (SD)	86.27 (19.82)	88.30 (17.41)

Continuation of Table C.3

	Historical sample	New random trial sample
median	84.90	86.30
min; max	36.30; 171.70	50.98; 145.33
change from baseline to week 26 HbA1c		
N	2642	399
mean (SD)	-1.48 (1.01)	-1.81 (1.01)
median	-1.40	-1.70
min; max	-5.30; 2.60	-6.20 ; 1.20
creatinine (umol/L)		
N	3308	419
mean (SD)	74.14 (18.56)	73.60 (15.64)
median	72.00	72.00
min; max	23.00; 409.00	36.00; 121.00
diabetes duration (years)		
N	3311	419
mean (SD)	9.67 (6.36)	9.55 (6.26)
median	8.67	8.47
min; max	0.30; 49.65	0.44; 34.24
diastolic blood pressure (mmHg)		
N	3309	419
mean (SD)	78.85 (8.53)	79.12 (8.37)
median	80.00	80.00
min; max	47.00; 116.00	57.00; 109.00
eosinophils Blood (%)		
N	3284	419
mean (SD)	2.65 (2.35)	2.56 (2.06)
median	2.10	2.00
min; max	0.00; 43.70	0.00; 15.20
erythrocytes (10^{12}/L)		
N	3297	419
mean (SD)	4.68 (0.45)	5.08 (0.48)
median	4.70	5.00
min; max	3.10; 7.40	3.60; 7.50
fasting plasma glucose (mmol/L)		
N	3268	411
mean (SD)	9.81 (2.65)	9.55 (2.53)
median	9.50	9.20
min; max	2.70; 22.60	3.60; 29.20
haematocrit blood (%)		
N	3264	419
mean (SD)	42.49 (4.16)	45.28 (4.24)
median	42.50	45.50

Continuation of Table C.3

	Historical sample	New random trial sample
min; max	22.80; 60.50	31.20; 58.40
hemoglobin A1C at baseline (%)		
N	3311	419
mean (SD)	8.39 (0.92)	8.28 (1.01)
median	8.30	8.10
min; max	6.60; 12.80	6.50; 13.50
high density lipoprotein cholesterol (mmol/L)		
N	3277	410
mean (SD)	1.18 (0.33)	1.21 (0.35)
median	1.14	1.14
min; max	0.21; 3.99	0.31; 2.69
height (m)		
N	3311	419
mean (SD)	1.67 (0.10)	1.68 (0.09)
median	1.67	1.68
min; max	1.36; 2.03	1.43; 2.01
low density lipoprotein cholesterol (mmol/L)		
N	3270	409
mean (SD)	2.46 (0.94)	2.42 (1.01)
median	2.36	2.28
min; max	0.00; 6.73	0.10; 7.10
leukocytes (10^9/L)		
N	3297	419
mean (SD)	7.33 (1.93)	7.93 (2.04)
median	7.10	7.80
min; max	2.80; 17.60	3.60; 15.80
lymphocytes blood (%)		
N	3284	419
mean (SD)	30.00 (7.88)	29.39 (7.66)
median	29.70	28.70
min; max	4.60; 71.00	10.70; 55.10
monocytes blood (%)		
N	3284	419
mean (SD)	5.88 (2.19)	5.83 (2.30)
median	5.70	5.70
min; max	0.00; 21.70	0.50; 17.20
neutrophils blood (%)		
N	3284	419
mean (SD)	60.93 (8.68)	61.83 (9.00)
median	61.10	62.20
min; max	16.60; 91.60	25.20; 86.50

Continuation of Table C.3

	Historical sample	New random trial sample
potassium (mmol/L)		
N	3304	419
mean (SD)	4.48 (0.42)	4.53 (0.41)
median	4.49	4.50
min; max	3.10; 7.00	3.30; 6.50
pulse (beats/min)		
N	3310	419
mean (SD)	75.31 (10.00)	75.56 (9.34)
median	75.00	76.00
min; max	45.50; 118.00	52.00; 108.00
sodium (mmol/L)		
N	3303	419
mean (SD)	139.73 (2.81)	140.19 (2.44)
median	140.00	140.00
min; max	121.00; 154.00	132.00; 148.00
systolic blood pressure (mmHg)		
N	3309	419
mean (SD)	131.53 (14.40)	129.69 (13.69)
median	131.00	130.00
min; max	90.00; 200.00	96.00; 171.00
thrombocytes (10^9/L)		
N	3269	419
mean (SD)	240.18 (64.34)	244.27 (64.91)
median	233.00	242.00
min; max	13.00; 611.00	63.00; 477.00
total bilirubin (umol/L)		
N	3304	419
mean (SD)	8.10 (4.33)	8.12 (4.73)
median	7.00	7.00
min; max	0.00; 36.00	1.00; 33.00
total cholesterol (mmol/L)		
N	3284	410
mean (SD)	4.54 (1.13)	4.59 (1.28)
median	4.43	4.43
min; max	0.93; 13.93	2.02; 11.37
triglycerides (mmol/L)		
N	3279	410
mean (SD)	2.07 (1.78)	2.23 (2.36)
median	1.65	1.70
min; max	0.24; 34.25	0.38; 27.80

C.6 Summary of categorical baseline covariates of the case study

Appendix Table C.4: Summary of the continuous baseline covariates.

	Historical sample		**New random trial sample**	
	N	(%)	N	(%)
sample size	3311		419	
sex				
female	1480	(44.7)	173	(41.3)
male	1831	(55.3)	246	(58.7)
race				
Asian	820	(24.8)	65	(15.5)
Black or African American	183	(5.5)	< 5	
Other	67	(2.0)	< 5	
White	2241	(67.7)	346	(82.6)
smoking status				
current	241	(7.3)	53	(12.6)
never	910	(27.5)	249	(59.4)
previous	378	(11.4)	116	(27.7)
region				
Asia	748	(22.6)	57	(13.6)
Europe	1303	(39.4)	228	(54.4)
North America	1015	(30.7)	89	(21.2)
South Africa	91	(2.7)	0	
South America	154	(4.7)	45	(10.7)
ethnicity				
Hispanic or Latino	461	(13.9)	68	(16.2)
not Hispanic or Latino	2806	(84.7)	351	(83.8)
titration target				
3.9-5.0 mmol/l	410	(12.4)	0	
4.0-5.0 mmol/l	2236	(67.5)	419	(100.0)
4.4-7.2 mmol/l	665	(20.1)	0	
blinding				
double-blinded	122	(3.7)	0	
open-label	3189	(96.3)	419	(100.0)
Biguanides				
yes (continued in trial)	2873	(86.8)	369	(88.1)
yes (discontinued in trial)	248	(7.5)	27	(6.4)
no	190	(5.7)	23	(5.5)
Sulfonylureas				
yes (continued in trial)	436	(13.2)	< 5	
yes (discontinued in trial)	1485	(44.9)	0	
no	1390	(42.0)	> 414	

Continuation of Table C.4

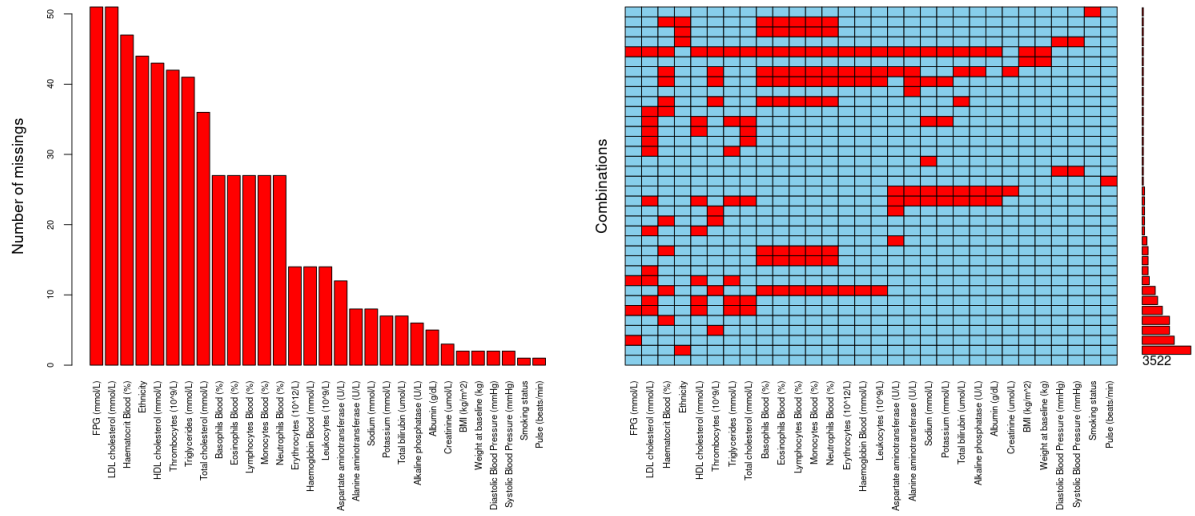
	Historical sample		**New random trial sample**	
	N	(%)	N	(%)
DPP4				
yes (continued in trial)	278	(8.4)	< 5	
yes (discontinued in trial)	361	(10.9)	> 123	
no	2672	(80.7)	> 285	
other blood glucose lowering drugs				
yes (continued in trial)	< 5		0	
yes (discontinued in trial)	> 79		0	
no	> 3220		419	(100.0)
Alpha Glucosidase inhibitor				
yes (continued in trial)	87	(2.6)	0	
yes (discontinued in trial)	69	(2.1)	0	
no	3155	(95.3)	419	(100.0)
combination of blood glucose lowering drug				
yes (continued in trial)	32	(1.0)	0	
yes (discontinued in trial)	36	(1.1)	8	(1.9)
no	3243	(97.9)	411	(98.1)
Thiazolidinediones				
yes (continued in trial)	78	(2.4)	20	(4.8)
yes (discontinued in trial)	29	(0.9)	0	
no	3204	(96.8)	399	(95.2)
SGLT2i				
yes (continued in trial)	175	(5.3)	> 383	
yes (discontinued in trial)	15	(0.5)	> 25	
no	3121	(94.3)	< 5	
GLP-1 receptor agonist				
yes (continued in trial)	79	(2.4)	0	
yes (discontinued in trial)	13	(0.4)	0	
no	3219	(97.2)	419	(100.0)

C.7 Missing pattern of the case study

To clean and curate the 14 data sets we imputed the HbA1C at week 26 value. For the historical sample the imputation was made using an ANCOVA model with last observed HbA1C measurement before landmark visit, time point of last measurement, baseline HbA1C, discontinuation prior to week 26 indicator and study-id as adjustment covariates. For the new trial data a similar approach was employed. However, in this case the last observed HbA1C measurement before landmark visit week 26, time point of last measurement, baseline HbA1C, discontinuation prior to week 26 indicator, region, treatment indicator and pre-study OADs were used as adjustment covariates. This was done in order to use a similar imputation as used in the original

analysis.

After imputing the HbA1C at week 26 value a total 94.4% of the participants had complete data for the combined historical and new trial data. The missingness of the covariates is displayed below. For the covariates we included missingness indicators and respectively imputed covariates using random forest [111]. This was done seperately on the historical and new trial data. The normalized root mean square error was 0.218 for continuous covariates and proportion of falsely classified is 0.004 for the historical data sample. The normalized root mean square error was 0.010 for continuous covariates and proportion of falsely classified is 0.023 for the new trial data. The missingness indicators of the historical sample did all overlap with the missingness indicators from NN9068-4229 trial. Since some of the covariates had near zero variance, were colinear or had large absolute correlation with each other we removed some of the covariates.



Appendix Figure C.1: Number of missing covariates (left) and combination of missingness of covariates (right).

Due to near zero variance, collinearity or high absolute correlation between the covariates, we excluded some of the values. Thus the baseline covariates used in the model where reduced to the following:

- age
- diabetes duration
- body mass index
- HbA1C
- height
- weight
- Alanine aminotransferase
- Albumin
- Alkaline phosphatase
- Aspartate aminotransferase
- Basophils
- Creatinine
- Eosinophils
- Erythrocytes

- fasting plasma glucose
- Haematocrit
- HDL cholesterol
- LDL cholesterol
- Leukocytes
- Lymphocytes
- Monocytes
- Potassium
- Sodium
- Thrombosytes
- Total bilirubin
- Total cholesterol
- Triglycerides
- Diastolic blood pressure
- pulse
- Systolic blood pressure
- country
- sex
- race
- smoking status
- region
- ethnicity
- Biguanides
- DPP4
- SGLT2I
- Previous OADs