This is an Accepted Manuscript of an article published by Taylor & Francis in *The American Statistician* on 8 Feb 2024, available at: https://doi.org/10.1080/00031305.2024.2303419

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

Lauren D. Liao[1,*], Yeyi Zhu[2], Amanda L. Ngo[2],
Rana F. Chehab[2], Samuel D. Pimentel[3]

[1]Division of Biostatistics, Berkeley, CA 94720
[2]Kaiser Permanente Northern California Division of Research,
Oakland, CA 94612
[3]Department of Statistics, Berkeley, CA 94720
[*]For correspondence email: *ldliao@berkeley.edu*

July 18, 2024

[*]The authors 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. Lauren D. Liao is a PhD candidate at Division of Biostatistics, Berkeley, CA 94720 (Corresponding Email: *ldliao@berkeley.edu*). Yeyi Zhu is a Research Scientist II (equivalent to associate professor) at Kaiser Permanente Northern California Division of Research, Oakland, CA 94612 and Associate Adjunct Professor at Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94143 (Email: *yeyi.zhu@kp.org*). Amanda L. Ngo is data reporting and analytics consultant III at Kaiser Permanente Northern California Division of Research, Oakland, CA 94612 (Email: *Amanda.L.Ngo@kp.org*). Rana F. Chehab is a research post doctoral fellow at Kaiser Permanente Northern California Division of Research, Oakland, CA 94612 (Email: *Rana.Chehab@kp.org*). Samuel D. Pimentel is an Assistant Professor at the Department of Statistics, University of California, Berkeley, Berkeley, CA 94720 (Email: *spi@berkeley.edu*).

Abstract

Observational studies of treatment effects require adjustment for confounding variables. However, causal inference methods typically cannot deliver perfect adjustment on all measured baseline variables, and there is often ambiguity about which variables should be prioritized. Standard prioritization methods based on treatment imbalance alone neglect variables' relationships with the outcome. We propose the joint variable importance plot to guide variable prioritization for observational studies. Since not all variables are equally relevant to the outcome, the plot adds outcome associations to quantify the potential confounding jointly with the standardized mean difference. To enhance comparisons on the plot between variables with different confounding relationships, we also derive and plot bias curves. Variable prioritization using the plot can produce recommended values for tuning parameters in many existing matching and weighting methods. We showcase the use of the joint variable importance plots in the design of a balance-constrained matched study to evaluate whether taking an antidiabetic medication, glyburide, increases the incidence of C-section delivery among pregnant individuals with gestational diabetes.

Keywords: Graphical Methods; Inference; Variable Selection.

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 (Zubizarreta et al., 2014), or weighting designs with high-variance weights that hurt precision (Miratrix et al., 2018). 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 (Athey et al., 2018), balance tolerances for weighting (Ben-Michael et al., 2021b), and covariate distances or balancing constraints for matching (Stuart, 2010; Pimentel et al., 2015; Bennett et al., 2020). 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 (Ahmed et al., 2006; Stuart et al., 2011; Greifer, 2021; Hansen and Bowers, 2008; Rosenbaum and Rubin, 1985). 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 (Sales et al., 2018; Ben-Michael et al., 2021a). 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 (Rosenbaum, 2005; Huang et al., 2023). 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 1 and signed measures shown in Supplemental 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.

2 METHOD

2.1 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 lead-

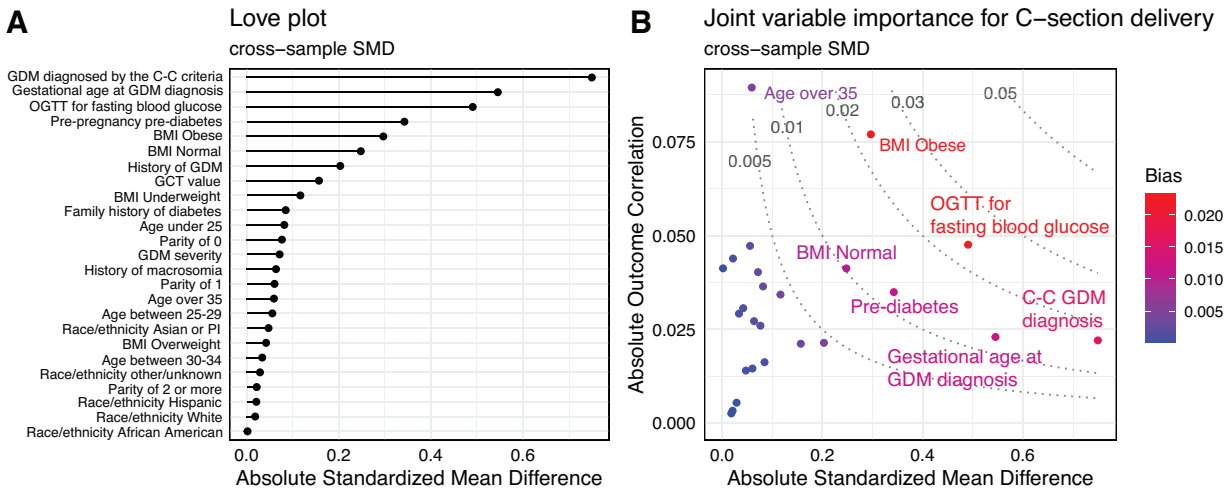

Figure 1: Comparison between the Love plot and the joint variable importance plot (jointVIP). Note that some variables (BMI in the obese category and OGTT for fasting blood glucose used at GDM diagnosis) take on much more prominent positions in jointVIP than in the Love plot, which only displays SMD values.
BMI: body mass index, C-C: Carpenter-Coustan, C-section: Cesarean section, GDM: gestational diabetes, OGTT: oral glucose tolerance test, PI: Pacific Islander, SD: standard deviation, SMD: standardized mean difference.

ing 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 (1976), which standardizes the mean difference by dividing by the standard deviation from the control group (Hedges, 1981). 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) (Pearson, 1895). 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 (2008) and Abadie et al. (2018).

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. (2020) 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.

2.2 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 (2020) and Soriano et al. (2021), 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:

$$Y = Z\tau_0 + X_j\beta_j + \hat{\epsilon} \tag{1}$$

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

$$Y = Z\tau + \hat{e} \tag{2}$$

$$X_j = Z\Delta_j + \hat{u} \tag{3}$$

Following Cochran's formula (Cox, 2007), we may use (3) to rewrite (1) and obtain a new representation for (2):

$$Y = X_j\beta_j + Z\tau_0 + \hat{\epsilon} = (Z\Delta_j + \hat{u})\beta_j + Z\tau_0 + \hat{\epsilon} = Z(\Delta_j\beta_j + \tau_0) + (\hat{u}\beta_j + \hat{\epsilon}) \tag{4}$$

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 (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

$$\tau - \tau_0 = \Delta_j \beta_j \tag{5}$$

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 (1) is fit on controls alone (as it will be in our pilot-sample approach), both (1) and (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.

$$\beta_j = r_{X_{j,pilot}, Y_{pilot}} \frac{S_{Y_{pilot}}}{S_{X_{j,pilot}}} \tag{6}$$

$$\Delta_j = \bar{X}_{j1,analysis} - \bar{X}_{j0,analysis} \tag{7}$$

Using (3), we obtain (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 (5) and rearranging, we obtain:

$$\frac{\Delta_j \beta_j}{S_{Y_{pilot}}} = r_{X_{j,pilot},Y_{pilot}} \frac{(\bar{X}_{j1,analysis} - \bar{X}_{j0,analysis})}{S_{X_{j,pilot}}} \tag{8}$$

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:

$$\text{cross-sample SMD} = \frac{\bar{X}_{j1,analysis} - \bar{X}_{j0,analysis}}{S_{X_{j,pilot}}} \tag{9}$$

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.

2.3 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 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:

$$\frac{1}{K}\sum_{k=1}^{K}[Y_{k1}(0) - Y_{k2}(0)](p_{k1} - p_{k2}) \tag{10}$$

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. (2018, §4) and Huang and Pimentel (2022). 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 (Guo and Rothenhäusler, 2023; Pimentel and Huang, 2023). 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 Supplemental Appendix A.2 for full derivation). Under similar assumptions, Rosenbaum (2005) 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.

2.4 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 (Hansen, 2004), or their marginal imbalance could be restricted via fine or refined balance constraints (Yang et al., 2012; Pimentel et al., 2015) as in our case study in Section 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 (Zubizarreta, 2015). 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 (Rosenbaum, 2002). Table 1 summarizes the process of creating and applying jointVIP for practitioners, and a simulation study in Supplemental 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 (2008) 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 (2008) 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. (2022).

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 repeat steps 3-5	for matching and weighting re-plot with post-design SMD repeat as desired

Table 1: Suggested procedure for use of the joint variable importance plot. As discussed in Section 2.1, the pilot sample typically consists of controls only. See Section 2.4 for further details on practical use of jointVIP.

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 (Pimentel et al., 2016). Even when all variables could be used for adjustment, it is wise to exclude to such variables since they can inflate unmeasured confounding bias (Brooks and Ohsfeldt, 2013; Ding et al., 2017). JointVIP enables 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 (Pimentel et al., 2016).

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.

3 CASE STUDY

3.1 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 (Castillo et al., 2014). The safety of glyburide, however, remains contentious due to potential transfer to the fetus through the placenta (American College of Obstetricians and Gynecologists, 2018). 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 summarizes selected baseline variables (see Supplemental Appendix A.4 for the full data summary). Missing values were imputed separately for each year using random forest (Stekhoven and Bühlmann, 2012). Details about the pattern of missing values and the imputation procedure are reported in Supplemental 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.

3.2 Design

3.2.1 Variable selection using jointVIP

JointVIP is constructed using the `jointVIP` package in R; for a brief software tutorial see Liao and Pimentel (2023). 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 (Pimentel and Huang, 2023). 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 bal-

Table 2: Summary of selected baseline variables for pregnant individuals with gestational diabetes.

BMI: body mass index, C-C: Carpenter-Coustan, GDM: gestational diabetes, OGTT: oral glucose tolerance test, SD: standard deviation A normal OGTT fasting blood glucose level is lower than 95 mg/dL. Abnormal indicates a result higher than normal.

		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)

ance 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 (Pimentel et al., 2015). 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 3 summarizes the chosen balance tiers for the design. Specific potential bias values can be found in Supplemental 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.

3.2.2 Matched Design

We conduct matching with refined covariate balance using the `rcbalance` package in R and the balance tiers in Table 3. Post-matched jointVIP results, reflecting new levels of balance after matching, are plotted in Figure 2.B. For variables that were specified, post-matched biases are compared to pre-matched biases in Supplemental 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

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 3: Balance tiers for refined covariate balance for each outcome, chosen using jointVIP plots.
BMI: body mass index, C-section: Cesarean section, GDM: gestational diabetes, jointVIP: joint treatment-outcome variable importance plot, OGTT: oral glucose tolerance test.

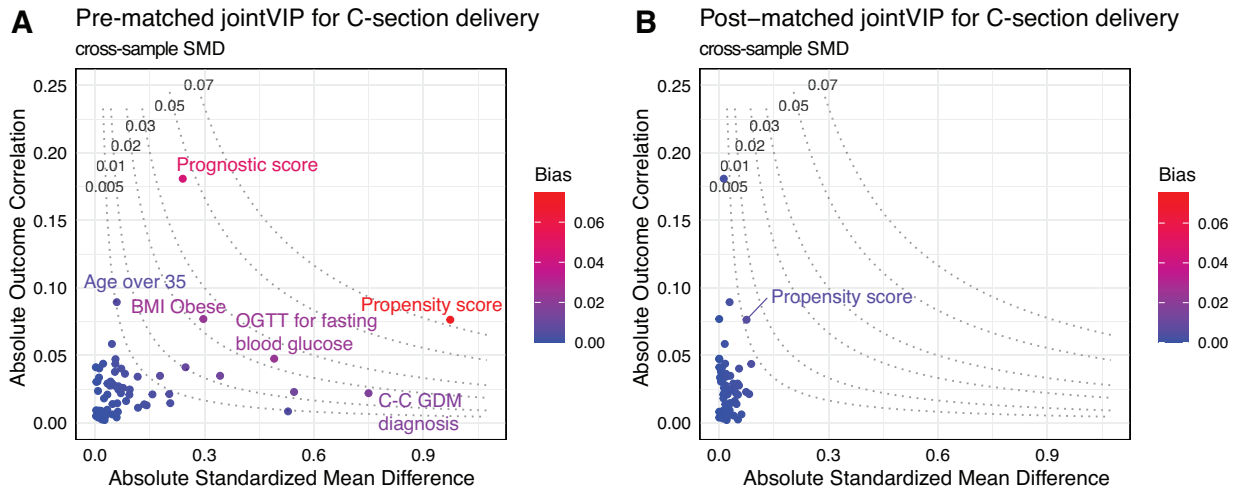

Figure 2: Pre-post match results for Cesarean section delivery.
BMI: body mass index, C-section: Cesarean section, GDM: gestational diabetes, jointVIP: joint treatment-outcome variable importance plot, OGTT: oral glucose tolerance test, SMD: standardized mean difference.

balanced especially well, a feature of the design that traditional methods based on Love plots are not equipped to guarantee.

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 (Supplemental 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 (Stuart, 2010), but as an average effect on a "marginal" population consisting of individuals for whom treatment by either arm is reasonably likely (Rosenbaum, 2012; Li et al., 2019; Greifer and Stuart, 2021). 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.

3.3 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 (McNemar, 1947). 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 (2022). 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 (2010) Section 3.

3.3.1 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 4. For control (MNT only) individuals, 33.61% delivered by C-section, and for treated (glyburide and MNT) individuals, 35.67% 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

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

Table 4: Matched analysis for Cesarean section delivery.

the increase in cases of C-section delivery in this study.

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 (Rosenbaum, 2005). Several other authors have discussed ideas closely related to jointVIP. Zhao and Yang (2022) 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. (2020) and Aikens and Baiocchi (2022) 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 (2020) 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 (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 (2020). 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 (2021) 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. (2023) 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.

Acknowledgments

The authors 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.

Conflict of interest

The authors report there are no competing interests to declare.

References

Abadie, A., Chingos, M. M., and West, M. R. (2018), "Endogenous stratification in randomized experiments," *Review of Economics and Statistics*, 100, 567–580.

Agarwal, A., Kenney, A. M., Tan, Y. S., Tang, T. M., and Yu, B. (2023), "MDI+: A Flexible Random Forest-Based Feature Importance Framework," *arXiv preprint arXiv:2307.01932*.

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

Aikens, R. C. and Baiocchi, M. (2022), "Assignment-Control Plots: A Visual Companion for Causal Inference Study Design," *The American Statistician*, 1–13.

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

American College of Obstetricians and Gynecologists (2018), "ACOG practice bulletin no. 190: gestational diabetes mellitus," *Obstet Gynecol*, 131, e49–e64.

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

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. (2021a), "The Balancing Act in Causal Inference," *arXiv preprint arXiv:2110.14831*.

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

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

Bicalho, C., Bouyamourn, A., and Dunning, T. (2022), "Conditional Balance Tests: Increasing Sensitivity and Specificity With Prognostic Covariates," *arXiv preprint arXiv:2205.10478*.

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

Castillo, W. C., Boggess, K., Stürmer, T., Brookhart, M. A., Benjamin Jr, D. K., and Funk, M. J. (2014), "Trends in glyburide compared with insulin use for gestational diabetes treatment in the United States, 2000-2011," *Obstetrics and gynecology*, 123, 1177.

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

Cox, D. (2007), "On a generalization of a result of WG Cochran," *Biometrika*, 94, 755–759.

Ding, P., VanderWeele, T., and Robins, J. M. (2017), "Instrumental variables as bias amplifiers with general outcome and confounding," *Biometrika*, 104, 291–302.

Glass, G. V. (1976), "Primary, secondary, and meta-analysis of research," *Educational researcher*, 5, 3–8.

Greifer, N. (2021), *cobalt: Covariate Balance Tables and Plots*, r package version 4.3.1.

Greifer, N. and Stuart, E. A. (2021), "Choosing the estimand when matching or weighting in observational studies," *arXiv preprint arXiv:2106.10577*.

Guo, K. and Rothenhäusler, D. (2023), "On the statistical role of inexact matching in observational studies," *Biometrika*, 110, 631–644.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

— (2008), "The prognostic analogue of the propensity score," *Biometrika*, 95, 481–488.

Hansen, B. B. and Bowers, J. (2008), "Covariate balance in simple, stratified and clustered comparative studies," *Statistical Science*, 219–236.

Hedges, L. V. (1981), "Distribution theory for Glass's estimator of effect size and related estimators," *journal of Educational Statistics*, 6, 107–128.

Huang, M. and Pimentel, S. D. (2022), "Variance-based sensitivity analysis for weighting estimators result in more informative bounds," *arXiv preprint arXiv:2208.01691*.

Huang, M., Soriano, D., and Pimentel, S. D. (2023), "Design Sensitivity and Its Implications for Weighted Observational Studies," *arXiv preprint arXiv:2307.00093*.

Li, F., Thomas, L. E., and Li, F. (2019), "Addressing extreme propensity scores via the overlap weights," *American journal of epidemiology*, 188, 250–257.

Liao, L. D. and Pimentel, S. D. (2023), "jointVIP: Prioritizing variables in observational study design with joint variable importance plot in R," *arXiv preprint arXiv:2302.10367*.

McNemar, Q. (1947), "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, 12, 153–157.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018), "Worth weighting? How to think about and use weights in survey experiments," *Political Analysis*, 26, 275–291.

Pearson, K. (1895), "VII. Note on regression and inheritance in the case of two parents," *proceedings of the royal society of London*, 58, 240–242.

Pimentel, S. D. and Huang, Y. (2023), "Covariate-adaptive randomization inference in matched designs," *arXiv preprint arXiv:2207.05019*.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "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, 515–527.

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

Rosenbaum, P. R. (2002), "Covariance adjustment in randomized experiments and observational studies," *Statistical Science*, 17, 286–327.

— (2005), "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *The American Statistician*, 59, 147–152.

— (2010), *Design of observational studies*, vol. 10, Springer.

— (2012), "Optimal matching of an optimally chosen subset in observational studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

Rosenbaum, P. R. and Rubin, D. B. (1985), "The bias due to incomplete matching," *Biometrics*, 103–116.

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

Soriano, D., Ben-Michael, E., Bickel, P. J., Feller, A., and Pimentel, S. D. (2021), "Interpretable sensitivity analysis for balancing weights," *arXiv preprint arXiv:2102.13218*.

Stekhoven, D. J. and Bühlmann, P. (2012), "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, 28, 112–118.

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

Stuart, E. A., King, G., Imai, K., and Ho, D. (2011), "MatchIt: nonparametric preprocessing for parametric causal inference," *Journal of statistical software*.

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

Zhang, Y. and Zhao, Q. (2022), "What is a randomization test?" *arXiv preprint arXiv:2203.10980*.

Zhao, A. and Ding, P. (2021), "Covariate-adjusted Fisher randomization tests for the average treatment effect," *Journal of Econometrics*, 225, 278–294.

Zhao, H. and Yang, S. (2022), "Outcome-adjusted balance measure for generalized propensity score model selection," *Journal of Statistical Planning and Inference*, 221, 188–200.

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

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014), "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, 204–231.

Supplementary Material for "Prioritizing Variables for Observational Study Design using the Joint Variable Importance Plot"

Lauren D Liao[1,*], Yeyi Zhu[2,3], Amanda L Ngo[2], Rana F Chehab[2] and Samuel D Pimentel[4]

[1]Division of Biostatistics, University of California, Berkeley, Berkeley, California, U.S.A

[2]Division of Research, Kaiser Permanente Northern California, Oakland, California, U.S.A

[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, U.S.A

[4]Department of Statistics, University of California, Berkeley, Berkeley, California, U.S.A

**corresponding email:* ldliao@berkeley.edu

Appendix A.1

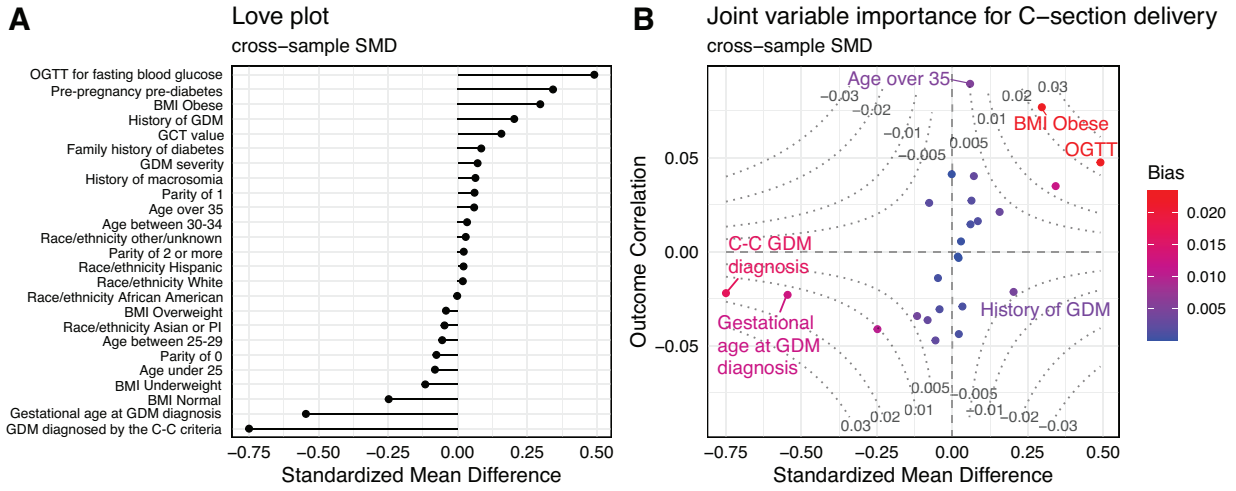

Figure A1. Comparison between the Love plot and the joint treatment-outcome variable importance plot with signed measures

NOTE: BMI: body mass index, C-C: Carpenter-Coustan, C-section: Cesarean section, GDM: gestational diabetes, jointVIP: joint treatment-outcome variable importance plot, OGTT: oral glucose tolerance test, PI: Pacific Islander, SD: standard deviation, SMD: standardized mean difference
Love plot (Supplemental Fig. 1.A) visualizing treatment imbalance alone using SMDs, and joint variable importance plot (Supplemental Fig. A1.B) visualizing both treatment imbalance and outcome model importance using SMDs and outcome correlations (among pilot controls) respectively. Signed measures are plotted on both axes.

Appendix 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:

$$Y(0) = \beta_0 X + \epsilon. \tag{11}$$

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 (Rosenbaum, 2010; Heng et al., 2021; Howard and Pimentel, 2021).

Consider the expected value of expression (10) over all realizations of $Y(0)s$ sampled from model (11). 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.

$$\begin{aligned}
&E_Y\left[\frac{1}{K}\sum_{k=1}^{K}[Y_{k1}(0) - Y_{k2}(0)](p_{k1} - p_{k2}) \,\middle|\, X_{11}, \ldots, X_{K2}\right] \\
&= \frac{1}{K}\sum_{k=1}^{K} E_Y[Y_{k1}(0) - Y_{k2}(0) \mid X_{11}, \ldots, X_{K2}](p_{k1} - p_{k2}) \\
&= \beta_0 \cdot \left[\frac{1}{K}\sum_{k=1}^{K}(X_{k1} - X_{k2})(p_{k1} - p_{k2})\right] \\
&= \beta_0 \cdot E_Z\left[\frac{1}{K}\sum_{k=1}^{K}(X_{k1} - X_{k2})(Z_{k1} - Z_{k2})\right] \\
&= \beta_0 \cdot E_Z\left[\bar{X}_{1,matched} - \bar{X}_{0,matched}\right]
\end{aligned} \tag{12}$$

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 (12) are parameters. However, under model (11) 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.

Appendix A.3

Simulation study setup

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.

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

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

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

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.

Design and estimation

In the simulated dataset, we conduct optimal pair matching using a Mahalanobis distance (Hansen and Klopfer, 2006). 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 (2007)).

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.

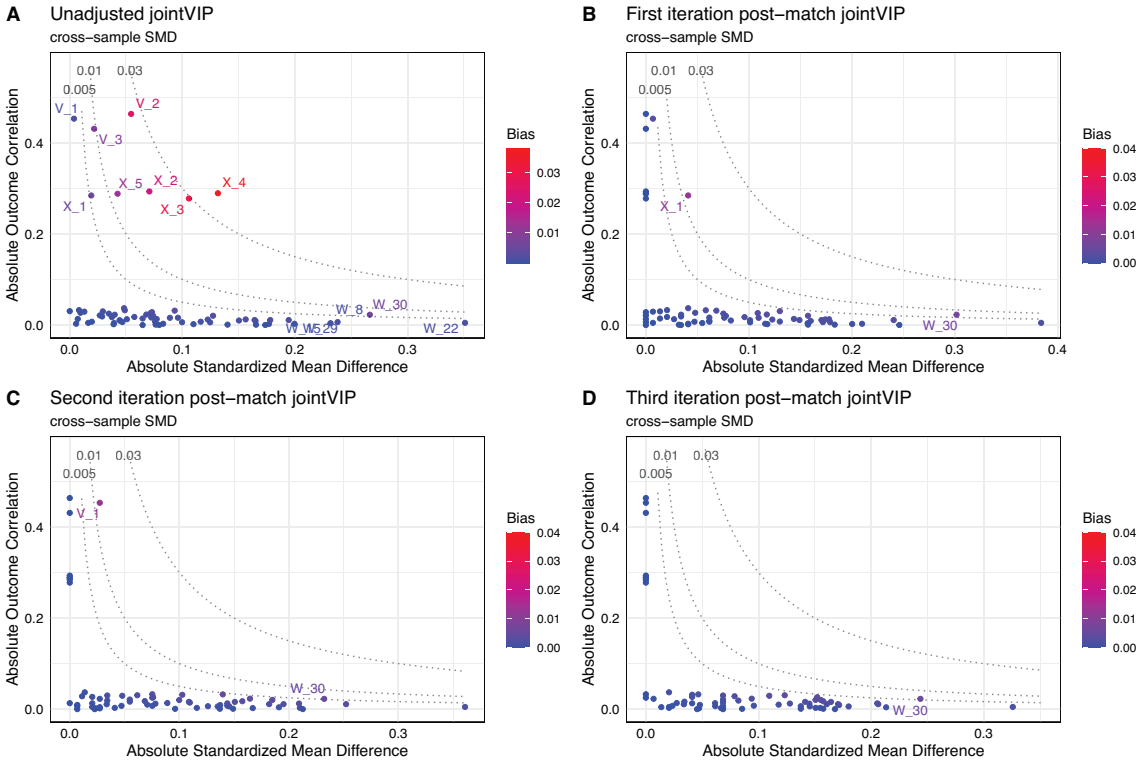

Figure A2. Iterative usage of the jointVIP showing all the variables are under 0.005 bias curve

For adjustment using jointVIP, the researcher would first examine the unadjusted jointVIP (Figure A2.A). The unadjusted plot indicates 6 variables needing adjustment above a 0.01 bias tolerance threshold. After the initial adjustment, the post-match jointVIP (Figure A2.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 (Figure A2.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 (Figure A2.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 A1. 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)

Table A1: Comparing different designs in the simulation, where the true treatment effect is 0.5.

Appendix A.4

Table A2: 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)
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)

Continuation of Table A2

		2007-2010 Control n = 7,526	2011-2021 Control n = 19,183	2011-2021 Treated n = 10,786
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)
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.

Appendix A.5

Table A3: 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).

Appendix A.6

Table A4: Pre-and-post matching comparison of background variables with high unadjusted bias.

C-section delivery

Background variable	Pre-matched bias	Post-matched 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-matched unadjusted bias for that outcome.

Appendix A.7

Table A5: Summary of all baseline variables of post-matched 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)
	$ 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)
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)

Continuation of Table A5

		C-section delivery included n = 8,693	C-section delivery excluded n = 2,093
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)

NOTE: 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.

References

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

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

Howard, S. R. and Pimentel, S. D. (2021), "The uniform general signed rank test and its design sensitivity," *Biometrika*, 108, 381–396.

Rosenbaum, P. R. (2007), "Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies," *Biometrics*, 63, 456–464.

— (2010), "Design sensitivity and efficiency in observational studies," *Journal of the American Statistical Association*, 105, 692–702.