Submitted to *Econometrica*

PROPENSITY SCORES SHOULD BE USED FOR MATCHING TO AVOID THE PARADOX OF SIMPLE DISTANCE METHODS

JASON MILLER

The use of Propensity Score Matching (PSM) was re-examined in light of the "Propensity Score Paradox" described by King and Nielsen (2018). Data from King's simulations were found to contain unreported statistical properties which mitigate the performance degradation of PSM in even the most highly artificial of conditions. Additionally, PSM was found to outperform Mahalanobis Distance Matching (MDM) when assumptions underlying the simulations' data generating process were altered to make them more reflective of real-world conditions in observational research.

KEYWORDS: propensity-score, matching, PSM.

1. INTRODUCTION

Propensity score matching (PSM) is a widely used quasi-experimental approach to inferring treatment effects based on observational data[1]. It is applied in many fields, notably healthcare and medicine, including the study of COVID-19, see: Biran, Ip, Ahn, Go, Wang, Mathura, Sinclaire, Bednarz, Marafelias, and Hansen (2020), Freedberg, Conigliaro, Wang, Tracey, Callahan, Abrams, Sobieszczyk, Markowitz, Gupta, and O'Donnell (2020), Geleris, Sun, Platt, Zucker, Baldwin, Hripcsak, Labella, Manson, Kubin, and Barr (2020), Goldman, Ye, and Scheinfeld (2020), Lee, Ha, Yeniova, Moon, Kim, Koh, Yang, Jeong, Moon, and Cho (2020), Liu, Lin, Baine, Wajnberg, Gumprecht, Rahman, Rodriguez, Tandon, Bassily-Marcus, and Bander (2020), Magagnoli, Narendran, Pereira, Cummings, Hardin, Sutton, and Ambati (2020), Mehta, Kalra, Nowacki, Anjewierden, Han, Bhat, Carmona-Rubio, Jacob, Procop, and Harrington (2020), Meng, Lu, Guo, Liu, Yang, Wu,

Jason Miller: jason@libreagora.com

[1] A count according to Google Scholar, accessed 11/30/2020 totaled 263,000 results, searching for: "propensity score" AND(matching OR matched OR match). This is a marked increased from only 2 years earlier when the same search conducted by King and Nielsen already totaled 127,000 articles.

Lin, Peng, Fu, and Li (2020), Ntaios, Michel, Georgiopoulos, Guo, Li, Xiong, Calleja, Ostos, González-Ortega, and Fuentes (2020), Paccoud, Tubach, Baptiste, Bleibtreu, Hajage, Monsel, Tebano, Boutolleau, Klement, and Godefroy (2020), Tremblay, van Gerwen, Alsen, Thibaud, Kessler, Venugopal, Makki, Qin, Dharmapuri, and Jun (2020), Zhang, Chen, Wang, Gong, Zhou, Cheng, Xia, Chen, Meng, and Ma (2020). Other areas of application include economics and law, see: Rubin (2001).

Key benefits of PSM include the ability to control for observable selection bias and that statistical modeling occurs only in the first of two stages, when the researcher estimates a model of the likelihood that a unit of observation was included in treatment based on a vector of covariates. This two-stage approach allows the researcher to avoid being biased from seeing the estimated treatment effect when selecting a model specification as would be the case with traditional regression methodologies – the propensity score (PS) model is estimated with "no outcome variables in sight", see: Rubin (2001).

Recently, some have proposed that PS models should not be used for matching, most notably King and Nielsen (2018). The basis for this claim was a hypothesized "PSM paradox", wherein the use of PS leads to higher sensitivity ("model dependence") and bias compared to using simple distance matching on covariates, such as by Mahalanobis distance matching (MDM). Since PSM involves creating a single vector index (the PS) of multiple covariates, PSM does not necessarily use all available information, as would matching on the raw values of all observed covariates with another Equal Percent Bias Reducing (EPBR) matching method such as MDM or a method of the Monotonic Imbalance Bounding (MIB) class that combines basic feature engineering (e.g. creating categorical variables out of continuous data or larger categorical variables[2]) with exact matching such as Coarsened Exact

[2]This is crucial to CEM and other exact matching approaches - except exact matching on the PS - because, as King and Nielsen (2018) noted, most data sets cannot be exactly matched on the raw features. Such "coarsening" necessarily results in a loss of information by ignoring potentially important variation, which is ironically a very analogous weakness to the potential information loss in creating a PS to match upon rather than matching directly raw data features. In a properly specified PS equation however, irrelevant variation can be ignored and the relative importance of covariates to treatment assignment is reflected in the magnitude of coefficients in the PS model. Without the first stage model of the PS to guide the compression of information, coarsening raw data features

Matching (CEM), see: Iacus, King, and Porro (2011) and Iacus, King, and Porro (2012). Simulations by King and Nielsen (2018) were purported to demonstrate that this use of a composite index leads to model dependence in PSM as pruning of the samples increased, while MDM had low variance and unbiased treatment effect estimates. Our re-examination of these "PSM paradox" simulations casts doubt on conclusions supporting the existence of any meaningful paradox. Assumptions underlying the data generating process (DGP) in these simulations were highly artificial, as were those in the analytical process (particularly the presumption that a researcher would review 512 model specifications to select the one closest to their desired result – an assumption of such extreme dishonesty that if it were fact, no statistical methodology would be able to overcome such zeal in fabricating results). Using the same simulated data sets and the same analytical process however, we found unreported summary statistics which indicated that PSM performed equally as well as MDM. Further, when we made even the most minor alterations to the DGP or analytical process to make them more reflective of real-world observational data studies the paradox failed to exist. When we made more moderate deviations PSM significantly outperformed MDM. We traced PSM's superior performance to its ability to incorporate statistical learning in its first stage that does not exist in other methods such as MDM or CEM and to its avoidance of using statistical models with the outcome variable for measuring treatment effects. The failure of MDM and similar methods to integrate this statistical learning safely apart from the estimation of treatment effects, has demonstrated potential to result in highly biased and sensitive estimates, which one might label a "Simple Distance Matching Paradox".

2. THE ILLUSION OF THE PSM PARADOX

2.1. *Origins of the Concept*

King and Nielsen (2018) described what they considered to be a "PSM paradox" with 2 supporting sets of simulations with comparison to matching based on simple distance calculations on raw features using MDM. The first simulation created a game where a data set

---

is more prone to losing important variation and retaining irrelevant information which leads to less efficient matching.

was generated using three DGPs – one representing a full block experiment, another representing a completely randomized experiment, and a third representing a disjoint subset of extra control units. There are several artificialities and questionable assumptions underlying the evidence for a "PSM paradox" as it pertains to observational data.

2.2. *The Inadequacy of Multiple Regression as an Identification Strategy for ATT and ATE*

Firstly, the simulations used to illustrate the supposed paradox did not carry out true (two-stage) PSM. In its first stage PSM is the estimation of a PS via logistic regression or other algorithm taking the functional form:

$$\alpha_x = pr(z = 1|x) \tag{2.1}$$

where α_x is the PS - a function of observable characteristics (x) that represents the probability that a unit was assigned to treatment ($z = 1$) given x. This is followed by a second stage estimation of causal (treatment) effects from matched pairs in the pruned sample with matching statistics such as Average Effect of Treatment on the Treated (ATT) or Average Treatment Effect (ATE).

Rather, the authors used PS and Mahalanobis distance only to rank matches for pruning then proceeded to perform Ordinary Least Squares (OLS) regression with the pruned samples to estimate treatment effects. We refer to this as Propensity Score Pruning OLS (PSP-OLS) and Mahalanobis Distance Pruning OLS (MDP-OLS) to distinguish it from actual PSM and MDM. The use of PSP-OLS defeats a major purpose of PSM, which is to avoid fitting models that estimate causal effects because doing so opens a door to bias in the model specification selection. Instead, ATT is calculated as:

$$ATT = \frac{1}{Nt} \sum_{i:Wt=1}^{N} (Yi(1) - Yi(0)) \tag{2.2}$$

where $Y_i(1)$ is the outcome for the unit i exposed to treatment ($W_{t=1}$) and $Y_i(0)$ is the outcome of a control unit matched to unit i based on observable characteristics. ATE is essentially the same measure, but for all units in the sample including controls. Only simple

(univariate) OLS regression yields treatment coefficients equivalent to ATT. Consider that simple linear regression may be expressed as:

$$Y = x\beta + \epsilon \tag{2.3}$$

where $\beta = (\alpha, \beta_1)^T$, ϵ is an error term and x is an $n \times 2$ matrix with observations as rows.

Assuming that the data, x, represents a binary vector of treatment assignment then the simple linear regression coefficient is equivalent to the mean of the dependent variable for treated units (treatment = 1) minus those for control units (treatment = 0):

$$E(y \mid x) = \alpha + \beta x, \tag{2.4}$$

$$E(y \mid x = 0) = \alpha, \tag{2.5}$$

$$E(y \mid x = 1) = \alpha + \beta, \tag{2.6}$$

$$E(y \mid x = 1) - E(y \mid x = 0) = \beta \tag{2.7}$$

In such a situation there is also only 1 possible model specification – the outcome regressed on treatment assignment, thus model dependency is irrelevant as human bias cannot enter the model specification selection.

When multiple linear regression or other multivariate models are used in the second stage, however, controls for additional covariates usually affect the magnitude of the treatment coefficient and the estimated quantity is thus no longer equivalent to ATT. Using matrix notation we can say that:

$$\hat{\beta}_{OLS} = (x^T x)^{-1} x^T y \tag{2.8}$$

for simple linear regression. Now, assume that we add a new vector of observations, x_a. Then, we can define a new design matrix $x_{new} = \begin{pmatrix} x & x_a \end{pmatrix}$ which is $n \times 3$ (allowing a column of ones for the intercept). Our new regression coefficients are:

$$\hat{\beta}_{new} = (x_{new}^T x_{new})^{-1} x_{new}^T y = \begin{pmatrix} x^T x \; x^T x_a \\ x_a^T x \; x_a^T x_a \end{pmatrix}^{-1} \begin{pmatrix} x^T y \\ x_a^T y \end{pmatrix} \tag{2.9}$$

where we're now estimating three coefficients in the vector $\hat{\beta}_{new}$, namely α (the intercept), β_1, and β_2. The cross terms $x^T x_a$ and $x_a^T x$ are responsible for why the estimates for α and β_1 change from $\hat{\beta}$ to $\hat{\beta}_{new}$ when all covariates are not perfectly orthogonal, as when working with real data. This was the case in 511 of 512 of King and Nielsen's models for each of their 100 simulated data sets and 85 pruning levels in the main simulation used to establish evidence for a "PSM paradox". By introducing these alternate specifications the door is opened to bias and model dependency.

This defeats a primary value proposition in the design of PSM – that treatment effects are calculated with no outcome variables in sight during modeling (which should only occur in the first stage). We ask the question – is the introduction of an obviously problematic deviation from PSM and its fallout reason to condemn the use of PS for matching, or simply evidence that PSM should be carried out as it was designed and according to best practices? We offer 5 simulations to help answer.

2.3. *Observational Data is Rarely Simple and Well-behaved*

Secondly, the assumptions underlying the DGP in both of King and Nielsen's simulations were extremely poorly reflective of reality. In their first simulation, it was assumed that observational data essentially consists of perfectly randomized experiments, full block experiments, and some dissimilar control subjects to be pruned as being clearly dissimilar. The authors then tasked their preferred matching method (MDM) and PSM with the task of discriminating between these three groups in the pruning process.

They then showed that MDM pruned the dissimilar control subjects first, followed by the pairs generated to simulate a fully randomized sample, and finally retained the full block experiment data. PSM pruned the dissimilar control subjects first, but showed less separation between the fully randomized and full block pairs. While this task was perfectly constructed to suit the MDM method, it is unclear that a sizable proportion of real-world

observational studies contain subsets of perfectly matched treatment and control pairs who do not differ from the larger sample systematically in terms of either treatment effects or selection bias.

More often perfect full block experiments do not exist within larger observational data sets and the task of the researcher instead is to filter out noise from the data and give a reasonable estimate of causal effects of an intervention on the population of interest. Even when heavy pruning can be used to identify a small subset of data approximating such perfect conditions, if this subset is systematically dissimilar to the population of interest then estimates based upon them will be biased (and simply re-labeling the estimated effect as being a "feasible" version does not provide value to decision makers interested in the larger group). It is more realistic to assume that data is sampled from the population of interest and while some reasonable amount of pruning occurs to remove noise and outliers, that the causal inference should be reflective of that sampled population.

2.4. *First-stage PS models are informative*

Thirdly, it was assumed that researchers estimate an arbitrary and fixed PS equation rather than learning about the relationship between covariates and treatment during this first stage of PSP-OLS (i.e. that specification selection does not occur in the first stage where it is appropriate by design, thanks to the dependent variable not being treatment rather than the outcome, but that a specification search occurs only in the second stage where it is inappropriate, which is what PSM is designed to avoid). This side steps a potential benefit of PSM and PSP-OLS over simple distance matching based methods like MDP-OLS, as it ignores information learned in the first stage model.

In a second simulation of simulated 100 data sets, 85 pruning caliper levels and 512 OLS regression model specifications, it was assumed that researchers will fit hundreds of model specifications and choose that which gives a treatment effect closest to their desired result (the maximum estimator of all 512 specifications). Two random covariates were generated from a uniform distribution. The outcome variable was a function of both covariates (which had equal influence on the outcome), treatment, and random noise. The effect of treatment

on the outcome was defined to be equal among all treated subjects – Y was drawn from a random normal distribution with $\mu = a + b + 2 * treat$ and $\sigma = 1$, where "treat" is a vector that equals 1 if treatment was received and 0 otherwise.

The DGP is well suited for simple distance matching (e.g. MDM) because the variables have equal importance in the outcome, thus their distances matter equally. The low dimensionality of the data set is also favorable to simple distance matching or exact matching methods, whereas PS offer the ability to compress higher dimensional data into a single composite index for matching where exact matching on covariates is infeasible and to weight these covariates according to their association with treatment assignment, thereby controlling for observed selection bias. A more DGP more reflective of real-world observational data might simulate selection bias, unequal influence of covariates on the outcome of interest, or allow treatment effects to vary but subsets of the sampled population. Model dependence was then measured as the variation across these 100 data sets and 512 specifications. Bias was measured as the extent to which the maximum estimator of the 512 specifications varied from the true treatment effect. PSM was shown to having increasing variance and bias when too much pruning occurred (the paradox) and it was pointed out that at certain pruning calipers PSM retained matched pairs which lacked common support in the covariates – but this is a weak point as enforcement of common support or more strict matching calipers are a well-known best practice for PSM and other matching methods, which is implemented in popular matching software packages (cite references). The same assumptions were made as in the first simulation regarding the PS model – that a researcher would estimate a fixed and arbitrary first stage model taking advantage of the unique ability to learn about the relationship between covariates and treatment in that stage, instead doing specification selection (only) in the second stage.

2.5. *Inspection of the Evidence*

We inspected the data from the 100 original simulated data sets as well as the 85 pruned sub-samples per matching method and 512 model specifications for PSP-OLS and MDP-OLS. We found that although the PSP-OLS estimates showed increasing variance when

over-pruned and a high maximum estimator of the 512 model specifications, the Mean Square Errors (MSE) of all 512 PSP-OLS models were consistently as small as those for MDP-OLS even with this artificial DGP suited to simple distance matching on covariates (Figure 1).

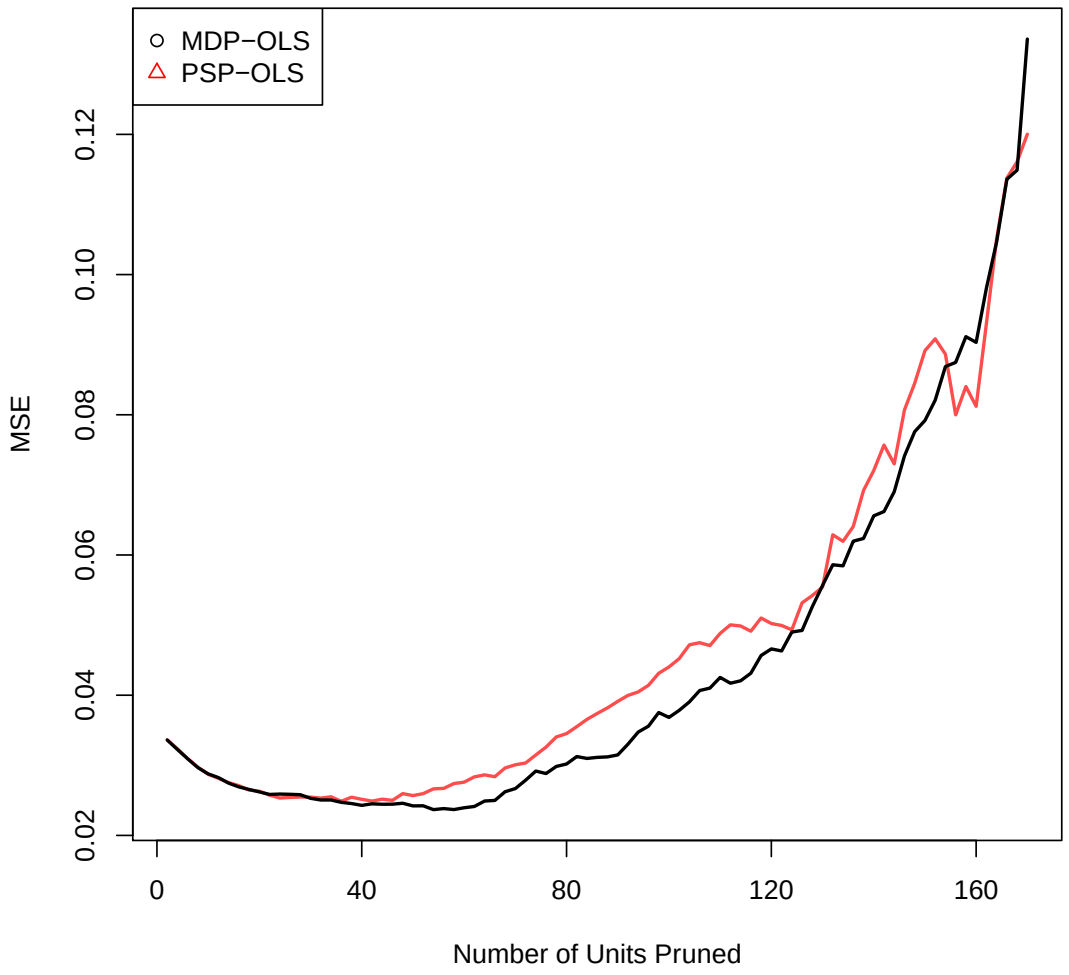

FIGURE 1.—MSE of the Mean Estimator of 512 Model Specifications across 100 Data Sets at 85 Pruning Calipers.

This would seem to imply that, even under the conditions of this simulation which was geared as a 'best case scenario' for simple distance matching on covariates, PSP-OLS still performed as well on average as MDP-OLS. For this reason an assumption of not slight bias, but extreme dishonesty – that an researcher would review as many as 512 model specifications and pick the most extreme estimate to support some desired or presupposed result – is necessary to illustrate failure of the PSP-OLS method relative to MDP-OLS.

2.6. *Simulation 1: PSP vs. MDP OLS with Selection Bias*

As a first step in checking the sensitivity of these results to the assumptions of the DGP and of extreme dishonesty by the modeler, we re-ran the simulation with a small number of alterations. We supposed that – while it is valid to assume that modelers may test many model specifications – instead of choosing the most extreme model of all 512 footnote: a degree of dishonest which, if present, cannot be overcome by any statistical means, as a researcher with such nefarious motives could ultimately always fake their results – whereas the potential for model-sensitivity under conditions reflective of an honest researcher seems more relevant a more realistic scenario would be that the modeler takes an average of the tested OLS specifications in the models of treatment effects on outcomes. We allowed the covariates to have unequal influence on the dependent variable by excluding one from the definition of Y and for one of the covariates to influence the probability of being assigned to treatment (selection bias).

Finally, we assumed that a modest amount of specification selection might occur in the first stage of PSP-OLS. We used stepwise regression on the raw covariates to simulate the process wherein a modeler may learn about the relationship between covariates and treatment then update their model accordingly. Although the assumptions more reflective of real-world observational data, the overall DGP was still somewhat favorable for simple distance matching on covariates, because the number of covariates was small, treatment effects were equal among all treated subjects, and PSP-OLS was used in place of PSM. The resulting estimates and MSEs were very comparable between PSP-OLS and MDP-

OLS, with PSP-OLS yielding slightly more accurate estimates of the treatment effects on average and roughly equal MSEs (Figure 2).

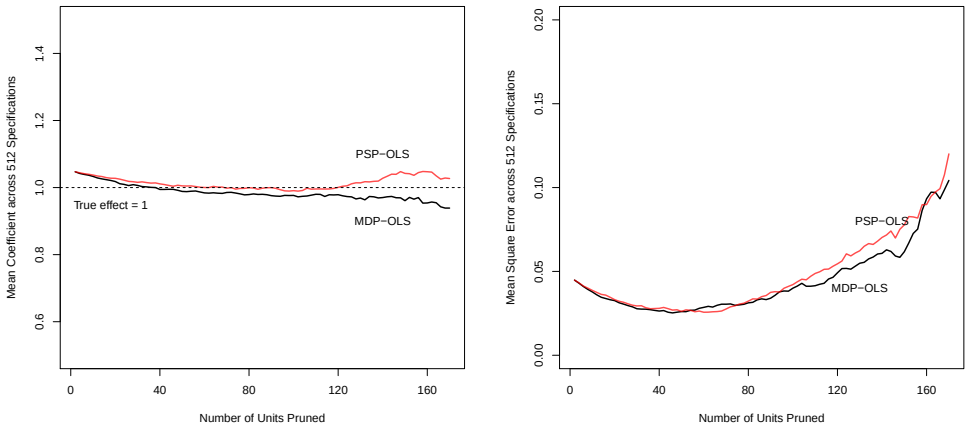

FIGURE 2.—PSP-OLS and MDP-OLS Mean Estimator (left) and MSE (right)

3. PROPER USE OF MATCHING ALGORITHMS IN OBSERVATIONAL RESEARCH

PSM is designed to be a two stage process where the first stage model of treatment assignment is estimated "with no outcome variables in sight" and the second stage calculation of treatment effects uses a model-free experimental statistics such as ATT or ATE (Rubin, 2001). Common best practices include limiting matching to areas of common support in the covariates and checking for balance of the covariates in the sample before and after pruning. Typically covariates that are imbalanced before pruning are including in the first stage PS model. Pruning occurs until there are no longer significant differences between treatment and control on the covariates. Given sufficient data, matching closeness calipers may be used to require closeness of matches beyond simply limiting to the regions of common support. Depending on sample size, the ratio of treatment to control, and contextual factors, the modeler may elect to perform many-to-one matching (many control units averaged per matching treatment unit) with or without replacement. Another common option is exact matching on the PS when there are sufficient matches to do so and there is no reason to believe that relevant segments of the sample will be excluded.

3.1. *Simulation 2: PSM vs. MDP-OLS with King and Nielsen's Data*

As a next step in our evaluation of the performance of PSM against simple distance matching alternatives, we compared PSM's estimation of ATT against treatment effects from MDP-OLS. We applied PSM and estimated ATT directly on King and Nielsen's original 100 data sets. We did not invoke any special matching options or best practices such as checking for balance, creating an ideal scenario for MDP-OLS (a friendly DGP and a minimalist implementation of the comparative matching method). Even with these ideal conditions, PSM had smaller errors than MDP-OLS at each of the 85 cutpoints and by virtue of the use ATT there was zero variance in the estimation of treatment effects (Figure 3).

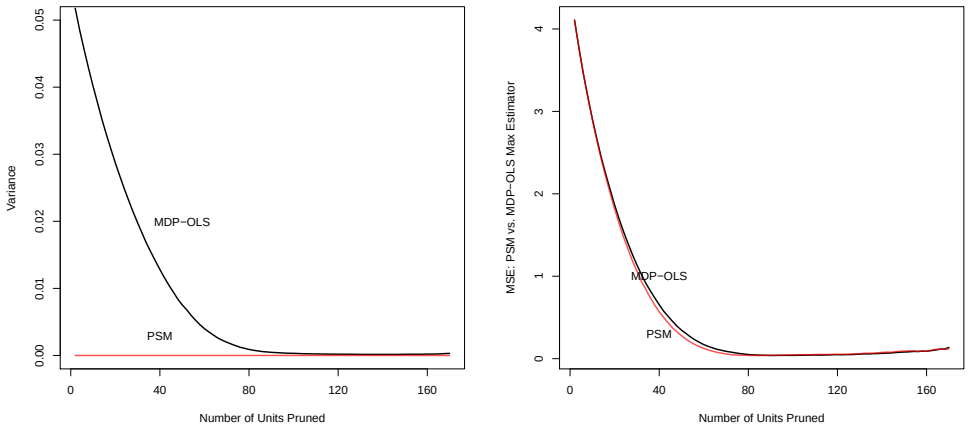

FIGURE 3.—PSM ATT vs. MDP-OLS Variance (left) and MSE (right) using King and Nielsen's data. The variance of PSM is constant at zero because there is no statistical model used to estimate treatment effects in PSM, precluding any potential for a "paradox" of increasing variance or bias from model dependence. PSM has comparable or lower MSE compared to the MDM max estimator using King and Nielsen's 100 simulated data sets at each pruning caliper.

3.2. *Simulation 3: PSM vs. MDP-OLS with Messy Data*

Continuing the along these lines, we next compared the PSM against MDP-OLS on 100 data sets and 85 pruning calipers generated with a slight alteration to the DGP to make it

more reflective of real observational data. We assumed that treatment effects might vary among segments of the sampled population, such as when a medication or program is more or less effective within different demographic groups. After a reasonable amount of pruning, the minimalist implementation of PSM produced an accurate and stable estimate of the treatment effect whereas MDP-OLS severely underestimated the effect (Figure 4). As pruning continued, the bias in MDP-OLS increased and although model dependency in MDP-OLS decreased in the middle pruning calipers, it increased towards the end. As always, PSM had zero variance in estimation of the treatment effects with ATT.

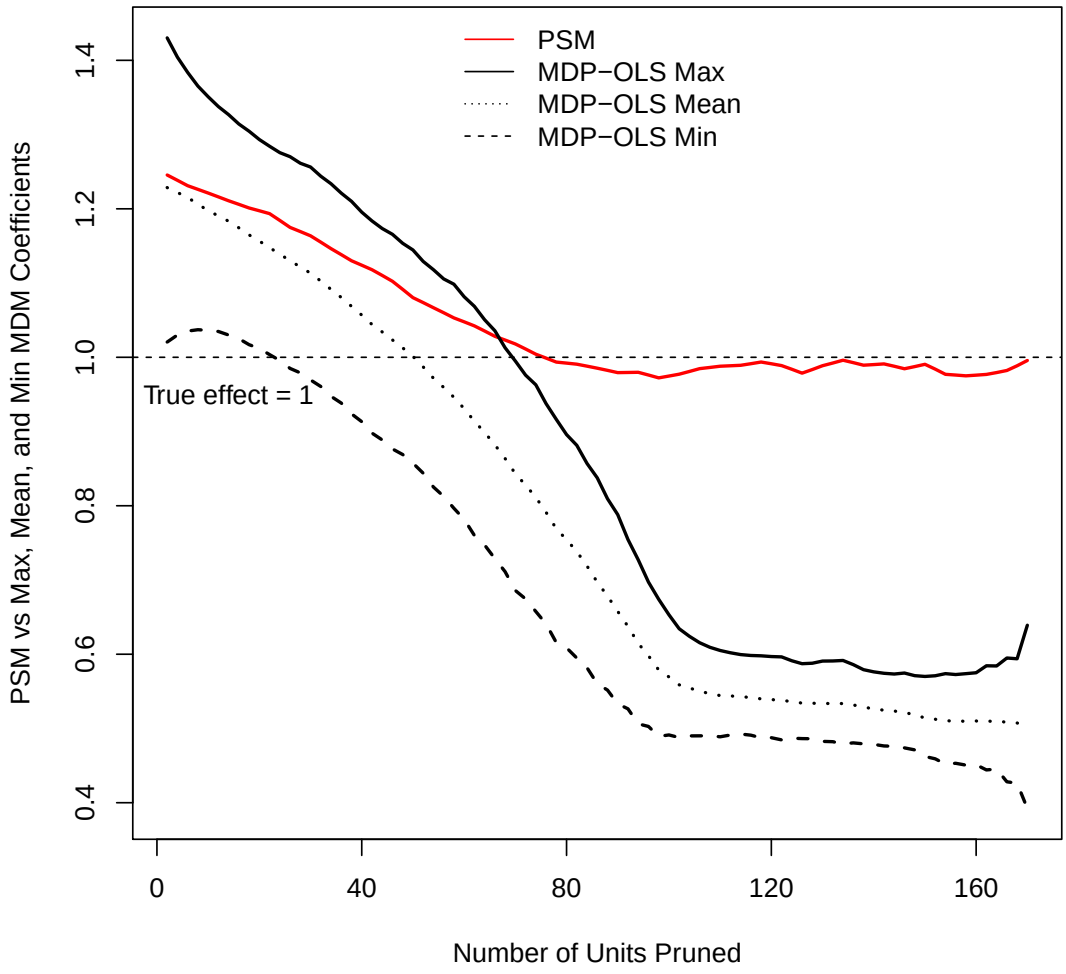

FIGURE 4.—PSM (red) vs. MDM Max, Mean, and Min Estimators using 100 simulated data sets with treatment effects varying by segment, 512 model specifications, and 85 pruning calipers. MDM underestimated the treatment effect because simple distance and exact matching methods cannot distinguish between important and unimportant matching features, causing a bias for selecting observations with arbitrary similarities at the expense of balance on features relevant to treatment assignment and outcomes - an ”exact matching paradox”. PSM correctly estimated the treatment effect by leveraging statistical information learning in the PS logistic regression step.

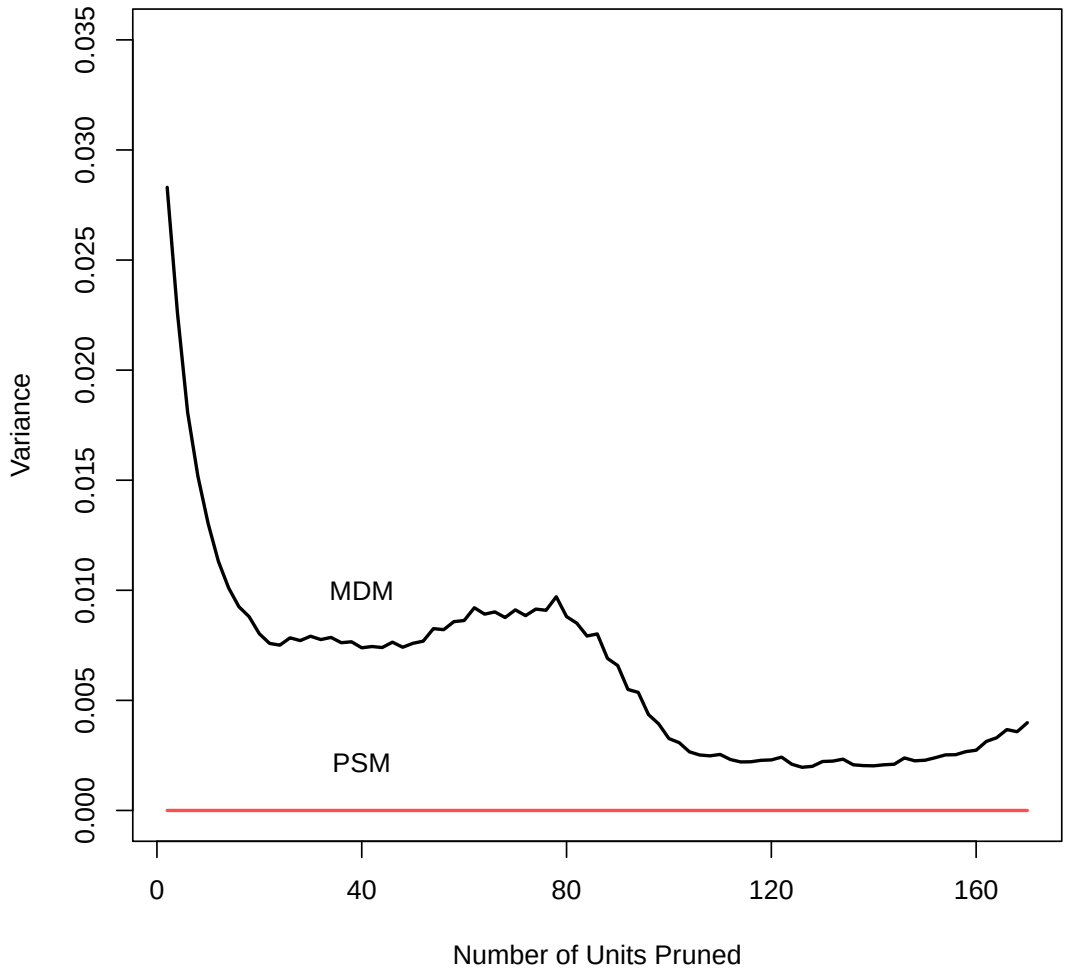

FIGURE 5.—PSM vs. MDM variance using the 100 new simulated data sets. PSM has no variance in estimating the treatment effects due to the use of ATT in place of OLS regression as in the MDM estimation strategy used by King and Nielsen (MDP-OLS).

3.3. *Simulation 4: PSM vs. MDM*

Finally, we compared the use of PSM with MDM in the estimation of ATT on two groups of 100 simulated data sets with 85 pruning calipers. The first group of data sets used the same simple DGP as in King and Nielsen's simulations, but defined the outcome as a ran-

dom sample of observations from a normal distribution with a mean of a + treatment (thus a true treatment effect of 1). The second group of data sets introduced selection bias and higher dimensionality (3 new variables) and correlations among the covariates. Both PSM and MDM performed well on the first group of data sets, with PSM generating only a slightly less biased causal inference (Figure 6 left). In the more realistic group of data sets PSM performed significantly better than MDM, however MDM did not suffer the high bias and model dependency as MDP-OLS did under similar conditions in the prior simulation (Figure 6 right).

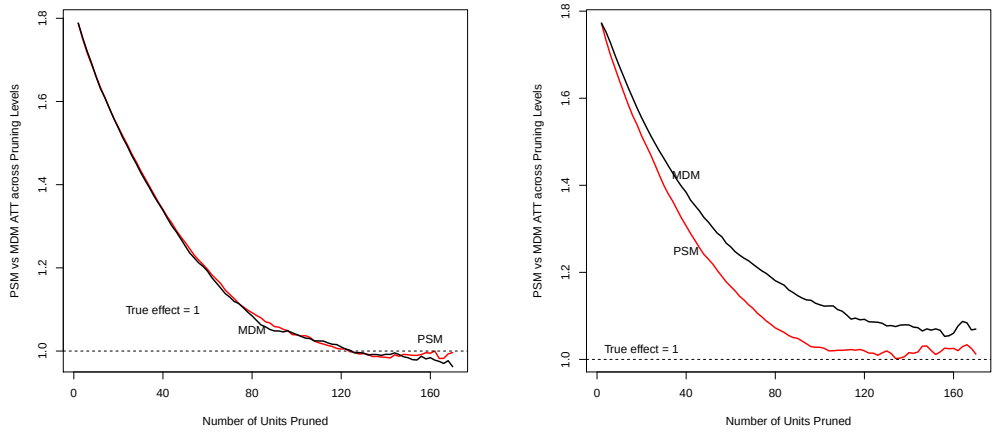

FIGURE 6.—ATT from PSM and MDM using data similar to that from King and Nielsen's DGP (left) and using new data that introduced selection bias and higher dimensionality (right).

4. PSM EXTENSIONS AND IMPROVEMENTS

Common best practices and options PSM include limiting matching to areas of common support in the covariates and checking for balance of the covariates in the sample before and after pruning. Typically covariates that are imbalanced before pruning are including in the first stage PS model. Pruning occurs until there are no longer significant differences between treatment and control on the covariates. Given sufficient data, matching closeness calipers may be used to require closeness of matches beyond simply limiting to the regions of common support. Depending on sample size, the ratio of treatment to control, and contextual factors, the modeler may elect to perform many-to-one matching (many control

units averaged per matching treatment unit) with or without replacement. Another common option is exact matching on the PS when there are sufficient matches to do so and there is no reason to believe that relevant segments of the sample will be excluded. These options are implemented and easily used in many matching software libraries such as "Matching" and "MatchIt" (see Sekhon (2008) and Blackwell, Iacus, King, and Porro (2009)). In addition, it has been demonstrated that various machine learning algorithms such as Random Forest can be used to estimate the PS in place of logit, often with greater model performance and increased automation, see: Austin (2014), Nichols and McBride (2019), and Goller, Lechner, Moczall, and Wolff (2020).

4.1. *Simulation 5: Matching Options*

As a test of how matching options may affect performance we ran a final simulation and applied PSM and MDM with and without common matching options. The DGP was the same as in Simulation 4, except that we once again allowed treatment effects to vary by population segment, as in Simulation 3. The estimates for both algorithms began much lower than the true treatment effect and improved rapidly, with PSM performing much better than MDM as pruning went on (Figure 7 top left) and the Mean Absolute Errors decreased accordingly (Figure 7 top right). The use of common matching options such as common support allowed the matching methods to much more quickly produce reasonable estimates (Figure 7 bottom left) and smaller MAEs (Figure 7 bottom right).

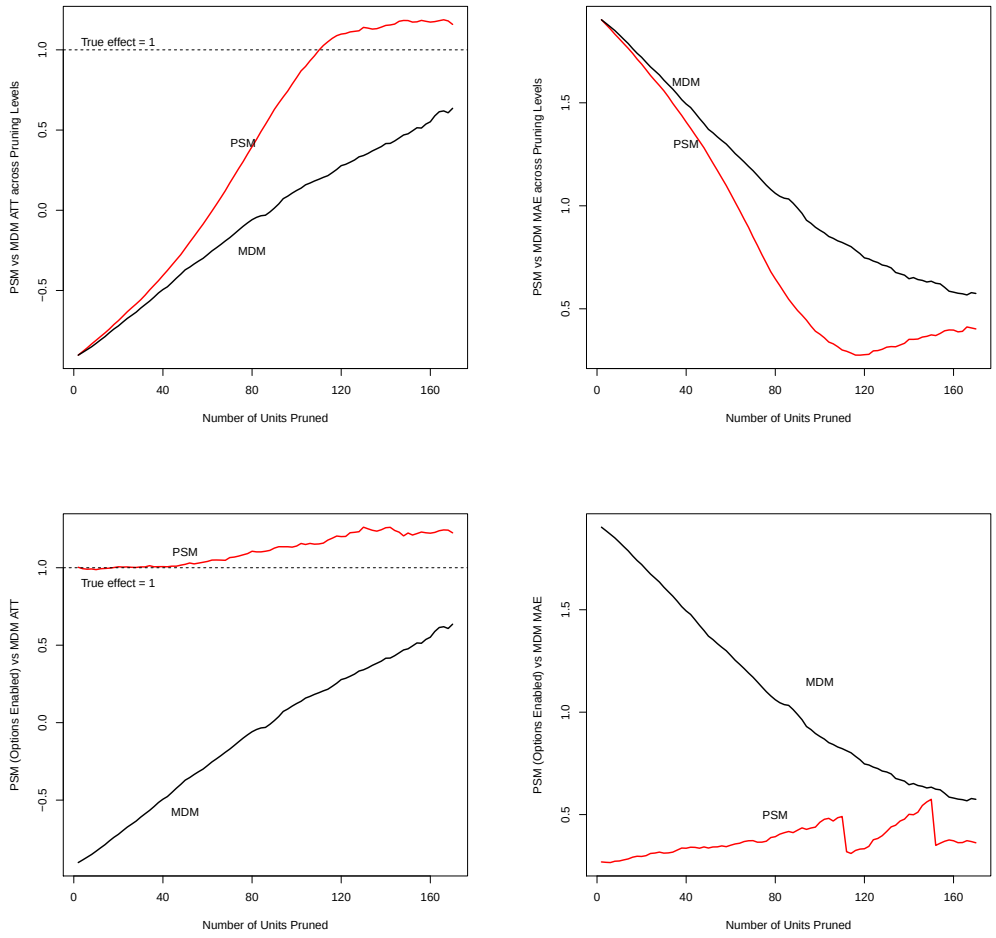

FIGURE 7.—PSM vs. MDM using the data set from Figure 6 but with differing treatment effects by segment. The top panels show ATT (left) and MAE (right) with no tuning options set. The bottom panels show ATT (left) and MAE (right) with basic tuning options enabled.

5. CONCLUDING REMARKS

The theme of reducing model dependence and bias in King and Nielsen (2018) was very appropriate, as it is of central importance to causal inference in observational studies[3]. The findings regarding use of PS for matching however cannot be generalized to applications of PSM that do not rely on a second stage regression model, to those implementing at least

[3] See also Rubin (2008)

some reasonable good practices, or to data sets that are of higher dimension and otherwise ”messier” data (i.e those that are not designed to work well with simple distance matching methods like MDM or MDP-OLS, as the simulated data sets used therein). Ironically, when faced with selection bias, unequal importance of covariates, varying treatment effect magnitudes or other complications reflective of the messy nature of observational data, MDM suffered from the same model dependence and bias that it was supposed to solve (a paradox of simple distance methods). Although PSM is far from a ”magic bullet” and requires that researchers follow good practices, the first stage statistical model provides a powerful tool unavailable in other methods which can be used to achieve better matching than other methods. Second stage regression models are not recommended due to their tendency to introduce bias and model dependency.

REFERENCES

AUSTIN, P. C. (2014): “A comparison of 12 algorithms for matching on the propensity score,” *Statistics in medicine*, 33(6), 1057–1069, ZSCC: 0000460 Publisher: Wiley Online Library.

BIRAN, N., A. IP, J. AHN, R. C. GO, S. WANG, S. MATHURA, B. A. SINCLAIRE, U. BEDNARZ, M. MARAFELIAS, AND E. HANSEN (2020): “Tocilizumab among patients with COVID-19 in the intensive care unit: a multicentre observational study,” *The Lancet Rheumatology*, ZSCC: 0000026 Publisher: Elsevier.

BLACKWELL, M., S. IACUS, G. KING, AND G. PORRO (2009): “cem: Coarsened exact matching in Stata,” *The Stata Journal*, 9(4), 524–546, Publisher: SAGE Publications Sage CA: Los Angeles, CA.

FREEDBERG, D. E., J. CONIGLIARO, T. C. WANG, K. J. TRACEY, M. V. CALLAHAN, J. A. ABRAMS, M. E. SOBIESZCZYK, D. D. MARKOWITZ, A. GUPTA, AND M. R. O’DONNELL (2020): “Famotidine use is associated with improved clinical outcomes in hospitalized COVID-19 patients: A propensity score matched retrospective cohort study,” *Gastroenterology*, ZSCC: NoCitationData[s0] Publisher: Elsevier.

GELERIS, J., Y. SUN, J. PLATT, J. ZUCKER, M. BALDWIN, G. HRIPCSAK, A. LABELLA, D. K. MANSON, C. KUBIN, AND R. G. BARR (2020): “Observational study of hydroxychloroquine in hospitalized patients with Covid-19,” *New England Journal of Medicine*, ZSCC: 0000680 Publisher: Mass Medical Soc.

GOLDMAN, I. A., K. YE, AND M. H. SCHEINFELD (2020): “Lower extremity arterial thrombosis associated with COVID-19 is characterized by greater thrombus burden and increased rate of amputation and death,” *Radiology*, p. 202348, ZSCC: 0000005 Publisher: Radiological Society of North America.

GOLLER, D., M. LECHNER, A. MOCZALL, AND J. WOLFF (2020): “Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany’s programmes for long term unemployed,” *Labour Economics*, p. 101855, Publisher: Elsevier.

IACUS, S. M., G. KING, AND G. PORRO (2011): "Multivariate matching methods that are monotonic imbalance bounding," *Journal of the American Statistical Association*, 106(493), 345–361, Publisher: Taylor & Francis.

——— (2012): "Causal inference without balance checking: Coarsened exact matching," *Political analysis*, pp. 1–24, Publisher: JSTOR.

KING, G., AND R. NIELSEN (2018): "Why propensity scores should not be used for matching," ZSCC: 0000641.

LEE, S. W., E. K. HA, A. Z. YENIOVA, S. Y. MOON, S. Y. KIM, H. Y. KOH, J. M. YANG, S. J. JEONG, S. J. MOON, AND J. Y. CHO (2020): "Severe clinical outcomes of COVID-19 associated with proton pump inhibitors: a nationwide cohort study with propensity score matching," *Gut*, ZSCC: NoCitationData[s0] Publisher: BMJ Publishing Group.

LIU, S. T., H.-M. LIN, I. BAINE, A. WAJNBERG, J. P. GUMPRECHT, F. RAHMAN, D. RODRIGUEZ, P. TANDON, A. BASSILY-MARCUS, AND J. BANDER (2020): "Convalescent plasma treatment of severe COVID-19: a propensity score–matched control study," *Nature medicine*, pp. 1–6, ZSCC: 0000007 Publisher: Nature Publishing Group.

MAGAGNOLI, J., S. NARENDRAN, F. PEREIRA, T. H. CUMMINGS, J. W. HARDIN, S. S. SUTTON, AND J. AMBATI (2020): "Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19," *Med*, ZSCC: 0000239 Publisher: Elsevier.

MEHTA, N., A. KALRA, A. S. NOWACKI, S. ANJEWIERDEN, Z. HAN, P. BHAT, A. E. CARMONA-RUBIO, M. JACOB, G. W. PROCOP, AND S. HARRINGTON (2020): "Association of use of angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers with testing positive for coronavirus disease 2019 (COVID-19)," *JAMA cardiology*, ZSCC: 0000132.

MENG, Y., W. LU, E. GUO, J. LIU, B. YANG, P. WU, S. LIN, T. PENG, Y. FU, AND F. LI (2020): "Cancer history is an independent risk factor for mortality in hospitalized COVID-19 patients: a propensity score-matched analysis," *Journal of Hematology & Oncology*, 13(1), 1–11, ZSCC: 0000015 Publisher: Springer.

NICHOLS, A., AND L. MCBRIDE (2019): "Propensity scores and causal inference using machine learning methods," pp. 21–23. ZSCC: 0000001.

NTAIOS, G., P. MICHEL, G. GEORGIOPOULOS, Y. GUO, W. LI, J. XIONG, P. CALLEJA, F. OSTOS, G. GONZÁLEZ-ORTEGA, AND B. FUENTES (2020): "Characteristics and outcomes in patients with COVID-19 and acute ischemic stroke: the global COVID-19 stroke registry," *Stroke*, 51(9), e254–e258, ZSCC: 0000013 Publisher: Am Heart Assoc.

PACCOUD, O., F. TUBACH, A. BAPTISTE, A. BLEIBTREU, D. HAJAGE, G. MONSEL, G. TEBANO, D. BOUTOLLEAU, E. KLEMENT, AND N. GODEFROY (2020): "Compassionate use of hydroxychloroquine in clinical practice for patients with mild to severe Covid-19 in a French university hospital," *Clinical Infectious Diseases*, ZSCC: NoCitationData[s0].

RUBIN, D. B. (2001): "Using propensity scores to help design observational studies: application to the tobacco litigation," *Health Services and Outcomes Research Methodology*, 2(3-4), 169–188, Publisher: Springer.

——— (2008): "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2(3), 808–840, Publisher: Institute of Mathematical Statistics.

SEKHON, J. S. (2008): "Multivariate and propensity score matching software with automated balance optimization: the matching package for R," *Journal of Statistical Software, Forthcoming*.

TREMBLAY, D., M. VAN GERWEN, M. ALSEN, S. THIBAUD, A. KESSLER, S. VENUGOPAL, I. MAKKI, Q. QIN, S. DHARMAPURI, AND T. JUN (2020): "Impact of anticoagulation prior to COVID-19 infection: a propensity score–matched cohort study," *Blood, The Journal of the American Society of Hematology*, 136(1), 144–147, ZSCC: NoCitationData[s0] Publisher: American Society of Hematology Washington, DC.

ZHANG, Y., R. CHEN, J. WANG, Y. GONG, Q. ZHOU, H.-H. CHENG, Z.-Y. XIA, X. CHEN, Q.-T. MENG, AND D. MA (2020): "Anaesthetic managment and clinical outcomes of parturients with COVID-19: a multicentre, retrospective, propensity score matched cohort study," *medRxiv*, ZSCC: NoCitationData[s0] Publisher: Cold Spring Harbor Laboratory Press.