PROFILE MATCHING IN OBSERVATIONAL STUDIES WITH MULTILEVEL DATA

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Brenda McGrath

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Statistics

May 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

PROFILE MATCHING IN OBSERVATIONAL STUDIES WITH MULTILEVEL DATA

By

Brenda McGrath

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Bong-Jin Choi
Chair

Dr. Megan Orr

Dr. Mingao Yuan

Dr. Pamela Jo Johnson

Approved:

5/25/2022
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Matching is a popular method to use with observational data to replicate desired features of a randomized control trial. A common problem encountered in observational studies is the lack of common support or the limited overlap of the covariate distributions across treatment groups. A new approach, cardinality matching, leverages mathematical optimization to directly balance observed covariates. When conducting cardinality matching, the user specifies the tolerable balance constraints of individual covariates and the desired number of matched controls. The algorithm then finds the largest possible match given these constraints. Profile matching is a newly proposed method that uses cardinality matching, in which the user can specify a target profile directly and find the largest cardinality match that is balanced to the target profile. We developed an R package called ProfileMatchit that will employ profile matching. We employed the new package in the setting of hospital quality assessment using a real-world dataset. Profile matching has not yet been used in hospital quality assessment but may be an improvement over current approaches, which have limitations in the ability to find sufficient matches in a heterogeneous sample. This application would be the culmination of our work to develop an improved version of cardinality matching and provide a new application of profile matching and a better approach to hospital quality assessment.

ACKNOWLEDGMENTS

I would like to first and foremost extend my gratitude to Dr. Bong-Jin Choi for accepting me as an advisee and guiding me to accomplish this dissertation project and throughout my doctoral degree experience. I especially appreciate that he always made time for me and genuinely cared about me personally and my success as a scholar. I would also like to thank my dissertation committee members Dr. Megan Orr, Dr. Mingao Yuan, and Dr. Pamela Jo Johnson for being part of this work. Finally, I acknowledge the Department of Statistics at North Dakota State University for the support and didactic experiences that they provided me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

APACHE.....Acute Physiology and Chronic Health Evaluation

APS .....Acute Physiology Score

ATE.....Average Treatment Effect

ATC.....Average Treatment Effect for the Control

ATT.....Average Treatment Effect for the Treated

CI.....Confidence Interval

O/E .....Observed/Expected

OR .....Odds Ratio

HS-TM .....Hospital-Specific Template Matching

ISM .....Indirect Standardization Matching

IS-PM.....Indirect Standardization Profile Matching

ICC.....Intraclass Correlation

IQR.....Interquartile Range

MI.....Multiple Imputation

MSE .....Mean Squared Error

RCM.....Rubin Causal Model

RCT.....Randomized Control Trial

SD .....Standard Deviation

SMD.....Standardized Mean Difference

SUTVA .....Stable Unit Treatment Value Assumption

TM.....Template Matching

LIST OF SYMBOLS

i .....Individual

j .....Cluster

m .....Covariate

N .....Number of individuals

C .....Number of control individuals

T .....Number of treated individuals

c.....Control individual

t.....Treated individual

p .....Number of covariates

k .....Balance constraint

$\boldsymbol{x}$ .....Covariate vector

$\boldsymbol{x_i}$ .....Observed covariate vector for individual i

x_{pi} .....Observed covariate p for individual i

x_{pij} .....Observed covariate p for individual i in cluster j

$\boldsymbol{X}$.....$n \times p$ covariate matrix

y_i .....Outcome for individual i

Z .....Binary indicator of treatment group

Z_i .....Treatment group indicator for individual i

$Z_i = 0$.....Indicates that individual i assigned to the control group

$Z_i = 1$.....Indicates that individual i assigned to the treated group

$\boldsymbol{z}$ .....Vector of treatment group indicators

$\hat{P}(\boldsymbol{x}_i)$ .....Estimated propensity score for individual i

β_p .....Coefficient for predictor p

$Y_i(Z_i)$ ..... Potential outcome for each individual i under each treatment assignment

$Y_i(0)$ ..... Potential outcome for each individual i under control treatment assignment

$Y_i(1)$ ..... Potential outcome for each individual i under treated treatment assignment

Y_i^{obs} ..... Outcome for individual i under treatment assignment that was observed

Y_i^{mis} ..... Outcome for individual i under treatment assignment that was not observed

$\hat{\mu}_0(x)$ ..... Estimation of the potential outcome under the control condition using linear regression

$\hat{\mu}_1(x)$ ..... Estimation of the potential outcome under the control condition using linear regression

$E[Y(0)]$ ..... Expectation of the outcome under the control condition

$E[Y(1)]$ ..... Expectation of the outcome under the treated condition

$\hat{E}[Y(0)]$ ..... Estimator of the expectation of the outcome under the control condition

$\hat{E}[Y(1)]$ ..... Estimator of the expectation of the outcome under the treated condition

$\boldsymbol{a}$ ..... Vector of matched pairs between the treated and control groups

a_{tc} ..... Indicator of whether treated individual t is matched to control individual c

$\boldsymbol{c}$ ..... Linear cost vector

$\boldsymbol{A}$ ..... Linear constraint matrix

$\boldsymbol{b}$ ..... Vector of linear constraints

K ..... Total number of balance constraints

$\mathbb{B}_k$ ..... Covariate balance constraint k

v_{ktc} .....Function of observed covariates

b_k .....Tolerance for the kth balance constraint

σ_j^2 .....Between-cluster variance

σ_w^2 .....Within-cluster variance

w_i .....Weight assigned to the individuals, equal to 1 for the treated individuals and the reciprocal of the number of matched controls for the control individuals

γ_j .....Cluster-level random intercept

ϵ_{ij} .....Residual error for individual i in cluster j

p_{ij} .....Probability of "success" for the outcome for individual i in cluster j

LIST OF APPENDIX TABLES

INTRODUCTION

The ideal experimental design to estimate treatment effects is a randomized control trial (RCT). In an RCT, the treated and control groups only randomly differ from one another on all covariates, both observed and unobserved[1]. Thus, any difference in outcomes can be attributed to the treatment rather than differences between individuals. However, an RCT is not always possible or feasible because of ethical considerations or the nature of the treatment.

Many studies rely on observational studies[2] when a randomized experiment is not possible or feasible. Observational studies are biased if there are differences in the treatment groups that matter for the outcomes under investigation[2]. This bias may be *overt* if the differences are observed or *hidden* if the differences are unobserved.

Matching can be used to estimate treatment effects with observational data when the treated and control groups have not been randomly allocated[3–7]. Treated and control individuals with similar covariates are compared, and thus matching replicates desired features of randomized experiments by creating groups that only randomly differ from one another on the observed covariates[1]. Matching highlights areas where there is insufficient overlap in the observed covariates between the treated and control groups.

When individuals are grouped in some way, for example patients within clinics or students within classrooms, the data structure is "multilevel" or "nested." With multilevel data, special consideration must be given to use matching techniques that properly account for the clustering[8]. Many-to-one matching can increase the precision of the treatment effect estimate in a single-level (i.e., non-nested) setting due to decreased variability, but also typically leads to an increased bias of the treatment effect because each additional matched control is less like the treated subject.

Cardinality matching[9] maximizes the size (or "cardinality") of the matched sample that satisfies the specified constraints for covariate balance. This approach uses recent advances in mathematical integer programming to solve an optimization problem subject to given constraints. The marginal distributions of the covariates in the treated and control groups are constrained to be similar, and the largest proportional match that keeps them similar is found. The algorithm is indifferent to which specific individuals are paired, so many solutions may achieve the same maximum cardinality.

Profile matching is a new multivariate matching method that finds the largest possible matched sample that is balanced relative to a reference covariate profile[10]. It is directly related to cardinality matching, which seeks to maximize the number of matched controls subject to balance constraints of the treated group. Profile matching seeks to maximize the size of both the treated group and the matched control group subject to the balance constraints of a provided profile.

The original implementation of cardinality matching was with the R package designmatch[11]. With designmatch, the user specifies the target and the tolerable difference from the target to the matched sample. For example, the user could specify the covariate means of the treated group as the target and set the tolerance to 0.05 standard deviations of the treated group. The treated and control groups would then be matched to the target; thus, the entire treated group would be retained. The means of the control group for each covariate would then be within 0.05 standard deviations of the treated group. In this example, the estimand is the average treatment effect for the treated (ATT)[1]. Alternatively, if the focal group of interest was the control, the estimand is the average treatment effect for the control (ATC).

Instead, the user could set the target to the be means of the entire sample (the treated and control groups combined), and the estimand would be the average treatment effect (ATE)[1]. The ability to specify the target directly makes the approach very flexible and in fact, the user could instead choose an external population to serve as the target to generalize inferences to another population, a method known as profile matching[10]. Additionally, the target is not limited to means. The user could also choose to impose balance constraints on higher order moments of the covariates. The limitation of designmatch is that it only finds a one-to-one match between the treatment groups.

The MatchIt[12] R package has also added cardinality matching to its suite of matching algorithms. It has the added benefit of allowing the user to specify many-to-one matching. Additionally, it can be used to find the largest possible subset that is balanced with respect to either the treated group, control group, or overall sample. For example, if the ATT is the estimand of interest, MatchIt can be used to select the largest possible control group that is balanced with respect to the treated group. The treated group remains intact by selecting the number of matched controls as infinity[12]. For the ATE, the largest possible sample from the full dataset that is balanced with respect to the combined treated and control group would be selected.

There are desirable features of the cardinality matching implementation from both the MatchIt and the designmatch packages. Thus, we created our own package, ProfileMatchit, which combines the desired features of each. Specifically, we have added the option to specify the target profile when conducting the cardinality matching method into the MatchIt package. The user can specify the desired target directly (as in designmatch) using our R package (ProfileMatchit) which can be used to conduct 1) traditional cardinality matching to estimate the

ATT, ATC, or ATE with one-to-one matching or many-to-one matching, 2) profile matching in which the user specifies a target profile directly, or 3) find the largest representative sample by setting the number of matched controls to infinity.

In brief, the user supplies a vector indicating the treatment group, a matrix of covariates, a target vector for each covariate, a vector of tolerable differences between the covariates and the target vector, the total number of matches (either a finite number for many-to-one matching or infinity to indicate largest possible subset), and the desired estimand (ATT, ATC, or ATE). The ProfileMatchit package then conducts mathematical optimization to find the largest possible matched sample that is balanced with respect to the target and tolerances given.

In this dissertation, we present the new R package called ProfileMatchit that will conduct many-to-one or largest subset profile matching. Then, we conducted a simulation study to examine the relative trade-offs of one-to-one matching, many-to-one matching, or largest subset selection and examined increasing the number of matched controls on bias, precision, accuracy, and coverage of the treatment effect estimate with multilevel data. We also compared a single-level linear model and a linear mixed effects model that accounts for the clustering of the control group to estimate the treatment effect in the matched sample. Then, we provide an example implementation of the ProfileMatchit package using real data in the context of hospital quality assessment.

BACKGROUND

Matching Methods

There are four main stages in matching: 1) defining closeness, i.e., which variables to include and the distance measure used to determine whether an individual is a good match for another; 2) implementing a matching method given the measure of closeness; 3) assessing the quality of the matched samples, and 4) analysis of the outcome and estimated treatment effect[1].

Propensity Score Matching

A commonly used distance measure for matching is the propensity score[13], which is the probability of assignment to the treated group, given a set of covariates. Matching on the propensity score yields balance in observed covariates between the treated and control groups on the distribution of the covariates. Differences in outcomes between treated and control individuals with similar propensity scores give unbiased estimates of the treatment effect. If treatment assignment is ignorable given the covariates, then the treatment assignment is also ignorable given the propensity score.

Propensity scores are most often estimated using logistic regression. With propensity score estimation, the resulting balance of the covariates is of interest rather than with the parameter estimates of the model. Thus, concerns with collinearity do not apply and standard approaches for model selection, such as model fit statistics identifying classification ability or stepwise selection models, are not useful for variable selection[14–16]. Studies have shown that misestimation of the propensity model, such as excluding a squared term of a covariate, is not as severe as misspecification of the outcome model[17–20].

With propensity score matching, individuals are matched on a one-dimensional propensity score so that the resulting matched pairs are heterogeneous in the covariates, but the

heterogeneity in the covariates is unrelated to the treatment, and thus tends to balance out in the treated and control groups. However, while randomization balances both observed covariates and unobserved covariates, matching only balances observed covariates[1]. The difference in the outcomes between the treated and control groups may then be due to differences in the treatment or may instead reflect some pretreatment differences in an unobserved covariate. Thus, matching can address overt biases but not necessarily hidden biases.

Many-to-One Matching

When there are many potential controls from which to match to a limited number of treated individuals, it may be desirable to match multiple controls to each treated individual, referred to as "many-to-one matching." With many-to-one matching, each treated individual is typically matched to one control, and then additional controls are added sequentially from the remaining controls (i.e., a second match is found for all treated individuals, then a third match is found for all treated individuals, and so on). This approach ensures that each treated individual is matched to its single best control individual.

Matching with more than one control can increase precision due to decreased variability of the estimated treatment effect[21, 22]. However, a trade-off is made because higher matching ratios typically increase bias in the estimated treatment effect because each additional matched control will be less similar to the treated subject, and fewer controls will be available later in the pool of matches[22]. The total number of controls that are matched may be either fixed ("fixed ratio matching") or may be allowed to vary ("variable ratio matching") so that each treated individual is matched to up to L controls. Ming and Rosenbaum showed that matching with a variable number of controls greatly reduced the bias compared to fixed ratio matching[23]. Studies have shown that matching ratios of up to 4-to-1 elicit the lowest bias in treatment

effect[21, 24, 25]. However, with cardinality matching, the user specifies *a priori* the tolerable differences between the treated and control groups, so concerns about balance are minimized because each matched control will be within the tolerable range on all covariates to be included.

Linden and Samuels, 2013 proposed the following approach[24] for selecting the optimal number of matched controls: 1) conduct the matching algorithm for one-to-one matching and iterating until the maximum number of desired potential controls per treated subject is reached, 2) for each iteration, test covariate balance, and 3) generate numeric summaries and graphical plots of the balance statistics across all iterations to determine the optimal solution. Austin (2010) conducted a Monte Carlo simulation to determine the optimal number of matched controls to estimate the treatment effect[22]. They varied the sample size, the proportion of the sample that was treated, the strength of the relationship between the observed covariates and the probability of treatment, and the strength of the relationship between observed covariates and the outcome[22]. The findings from this study indicated that increasing the number of matched controls increased the bias of the treatment effect but decreased the variance of the treatment effect. Thus, the authors recommended that studies using propensity score matching should use one or two matched controls for each treated subject[22].

Limited Overlap of Covariate Distributions

A common problem encountered in observational studies is the lack of common support or the limited overlap of the covariate distributions across treatment groups[26] which can lead to estimates that are sensitive to model misspecification[27, 28]. Overlap refers to the range of the data that is the same across treatment groups[29]. Complete overlap exists if the range of the data is the same between the treated and control groups[29]. When there is a lack of overlap between treatment groups, regression models rely on extrapolation[1, 29]. When the treated and

control groups do not completely overlap, regression models are inherently limited in the treatment effect estimation outside the region of overlap. The options are to either restrict the inferences to the regions of overlap or to rely on the model to extrapolate outside the region of overlap[29]. Importantly, the overlap is not the same as the imbalance. Imbalance does not necessarily imply a lack of complete overlap, nor does the lack of complete overlap imply imbalance[29].

Matching highlights areas of limited overlap. The traditional approach to handling the lack of covariate imbalance with matching is trimming the sample, such as discarding all individuals with estimated propensity scores outside the range of 0.1 and 0.9[28, 30]. If a large portion of the sample is lost after trimming regions of non-overlap, it could indicate insufficient overlap between the treated and control groups[31]. Cardinality matching[9] handles the setting of limited covariate overlap by directly balancing the original covariates and finding the maximum number of observations that satisfy any given covariate balancing criteria. With cardinality matching, the marginal distributions of the covariates in the treated and control groups are constrained to be similar, and the largest matched sample that keeps them similar is found.

Profile Matching

The goal of matching is to find similar individuals across the treatment groups to replicate a randomized experiment. Propensity score matching does not guarantee adequate balance between treatment groups. Because matching is generally conducted using the estimated propensity score (due to the true propensity score being unknown), propensity score model misspecification can yield problems[32]. If there is limited overlap in the distribution of the covariates between treatment groups, propensity score matching may require treated individuals

to be discarded[33] or important covariates to be excluded from the propensity score model, which limits the utility of the matching method.

To overcome these limitations, advances in matching methods have been made that leverage optimization. Examples of these methods include mixed integer programming[34], genetic matching[35], optimal matching with refined covariate balance[36], cardinality matching[9], and optimal matching using Glover's algorithm[37]. Profile matching is the latest advancement that solves an optimization problem that maximizes the sample size conditional on covariate balance restraints[10].

Profile matching is a new multivariate matching method that finds the largest possible matched sample that is balanced relative to a reference covariate profile[10]. It is directly related to cardinality matching and can be implemented using the existing designmatch[11] software in R. While cardinality matching seeks to maximize the cardinality of the matched controls subject to balance constraints of the treated group, profile matching seeks to maximize the cardinality of the matched controls subject to balance constraints of a provided profile.

Profile matching was introduced as a flexible approach that can aid in the generalization of causal inferences to a new target population or personalize causal inferences for an individual. Cohn and Zubizarreta illustrate this method in a simulation study that generalizes a randomized trial to a new target population, which may not have been well represented in the original trial[10]. This approach could be used for hospital performance comparisons; however, it has not yet been used in this context[10]. A conceptual diagram of profile matching in the context of matching patients in hospitals is presented in Figure 1.

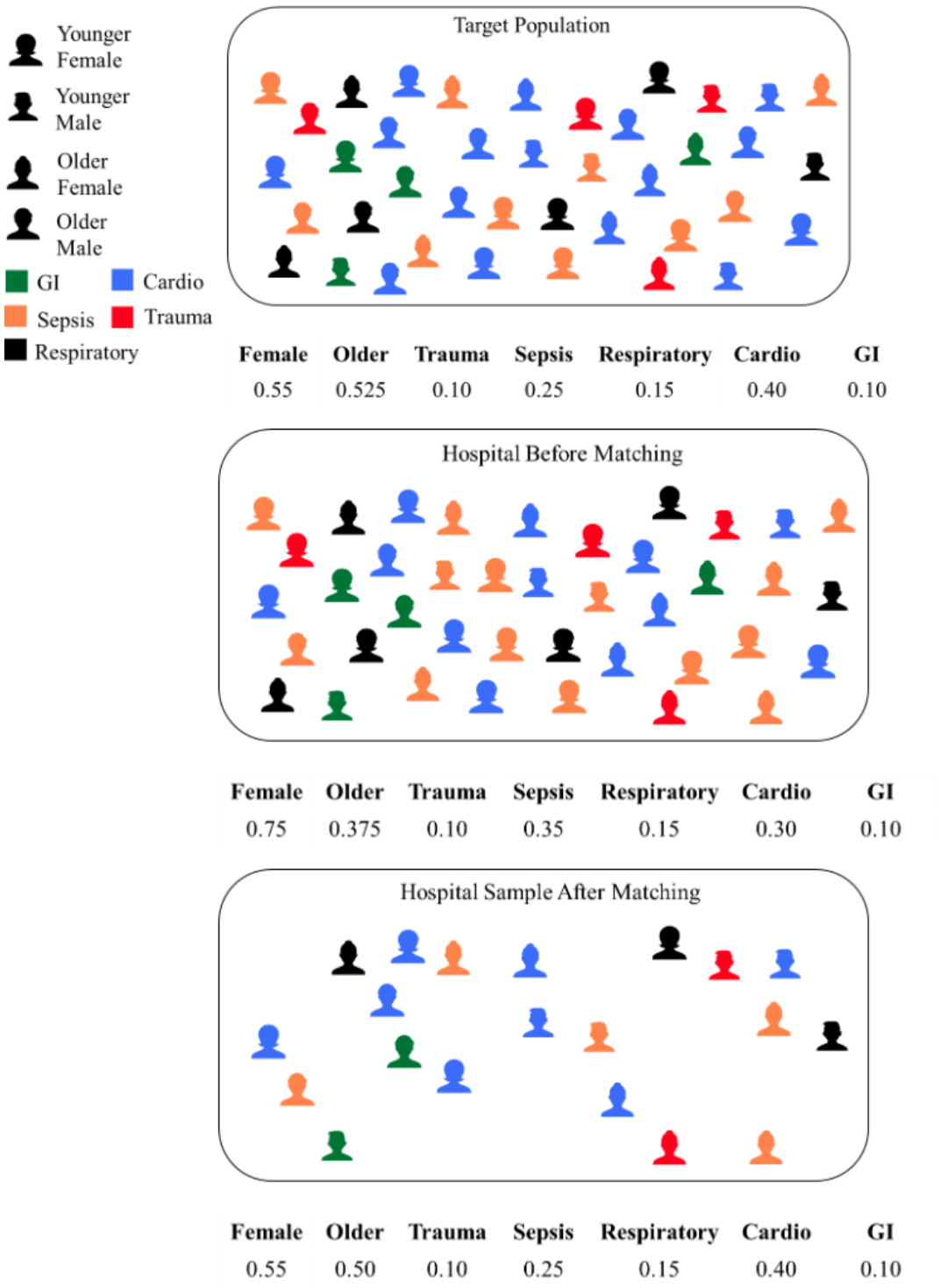


Figure 1. Diagram of Profile Matching
Note: A profile is established from the target population. In this example, we consider the proportion of females, the proportion of older age, and the proportions of five diagnosis groups. The hospital before matching had a notably different breakdown than the profile. Patients are selected from the hospital so that the overall population resembles the target.

Hospital Benchmarking

Patients are not randomly allocated to hospitals, and thus, hospital comparisons are limited by differences in patient case-mix and illness severity. Comparisons of patient outcomes between hospitals must consider these differences. Hospital performance assessment is traditionally done using regression models that adjust for differences in patient illness severity and other characteristics (e.g. age, diagnoses, comorbidities) to derive case-mix adjusted standardized mortality ratios so that comparisons between healthcare providers (such as a hospital) can be made[38, 39]. For example, the US Center for Medicare and Medicaid Services uses risk-adjustment to assign hospitals a star ranking by using average scores on five measure groups: mortality, the safety of care, readmission, patient experience, and timely and effective care[40]. There is some controversy about whether variation in standardized mortality ratios reflects differences in quality of care[41–45], but the practice is still in widespread use across hospital systems[39, 46–49] as part of their assessment of hospital quality.

The standardized mortality ratio is defined as the ratio of observed to expected deaths for a given hospital. It is derived from a statistical model that adjusts for the patient and/or hospitalization characteristics. A multilevel model that accounts for the nesting of patients in hospitals should be used to avoid the overestimation of systematic between-hospital effects[50–54]. However, a major limitation is that regression can yield biased estimates if there is insufficient overlap in patient covariates between the hospitals, as it relies heavily on extrapolation. In a heterogeneous hospital system, there may be limited overlap between the case-mix of patients between hospitals, and it is unlikely that any two hospitals have the same case-mix[55]. Additionally, multilevel regression models smooth the performance of low-volume hospitals towards the average (i.e., “shrinkage towards the mean”), precluding them from

being positive or negative outliers[56, 57]. To overcome these limitations, studies have examined the feasibility of matching in the context of hospital performance[58–64]. Matching methods have been used to compare outcomes between teaching and non-teaching hospitals[65, 66] and to compare the value of better and worse nursing environments[67]. Other methods have been proposed as a means of comparing hospital outcomes in a more general setting.

There are two ways to compare hospital outcomes after risk adjustment, either direct standardization or indirect standardization[68]. Direct standardization compares a hospital's patient outcomes to an external reference population and ultimately answers the question: "How does this hospital compare to other hospitals if all hospitals treated the same population of patients?". Indirect standardization instead examines the outcomes of a given hospital's patient population and assesses: "How would the outcomes of patients at a given hospital change had they instead gone to a different hospital that treats similar patients?".

Direct standardization and indirect standardization can yield different results. A hospital may have particularly good outcomes for some types of patients and poorer outcomes for other types of patients. A hospital that does not see the types of patient conditions that it cannot treat effectively may score well in indirect adjustment and poorly in direct adjustment. Conversely, a hospital that frequently sees the types of patient conditions that it cannot treat well may score poorly in both indirect and direct adjustment.

Indirect standardization approaches, such as the standard regression-based observed over expected (O/E) approach utilize the case-mix of the hospital's specific patients when estimating outcomes. Importantly, this approach relies on a model to appropriately adjust for each hospital's case mix. Model-based indirect standardization comparisons across hospitals require an adequate estimation of the "expected" number of deaths. Additionally, hospitals are compared on the

relative performance of their own case-mix to the other hospitals' relative performance on their case-mix, which may not be of clinical relevance[68]. A new method, called indirect standardization matching (ISM), was developed to address the limitations of a model-based indirect standardization approach using matching.

Studies have examined the feasibility of matching in the context of hospital performance[58–64]. The first approach, template matching[59], uses a common sample of patients to which all hospitals are matched. Another proposed approach is hospital-specific template matching[60], in which each hospital has a customized template, and hospitals are only compared to other hospitals that treat similar patients. Additional details of these approaches are provided in the next section.

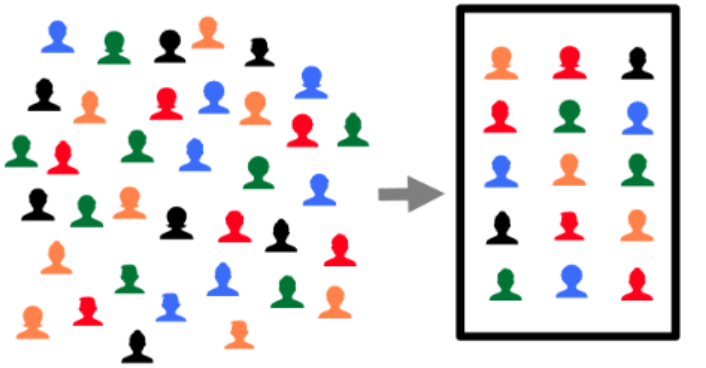
Template Matching

Template matching[59] uses a common sample of patients to which all hospitals are matched, and is therefore a form of direct standardization. With template matching, a randomly selected and representative sample of N hospitalizations from the population serves as the "template." At each hospital, one hospitalization is matched to each of the N hospitalizations in the template so that the resulting sample consists of N hospitalizations from each hospital. Hospitals can then be directly compared on patient outcomes, and hospitals are only being compared on the outcomes of similar patients. A conceptual diagram of TM is presented in Figure 2. Studies have found template matching feasible in a relatively limited population, such as within limited diagnoses[58, 62, 69] or surgical type[59], but fail in a more heterogeneous population[61, 63].

When this approach was originally introduced, Silber et al. (2014) illustrated the method using a limited set of Medicare patients admitted for orthopedics and common general,

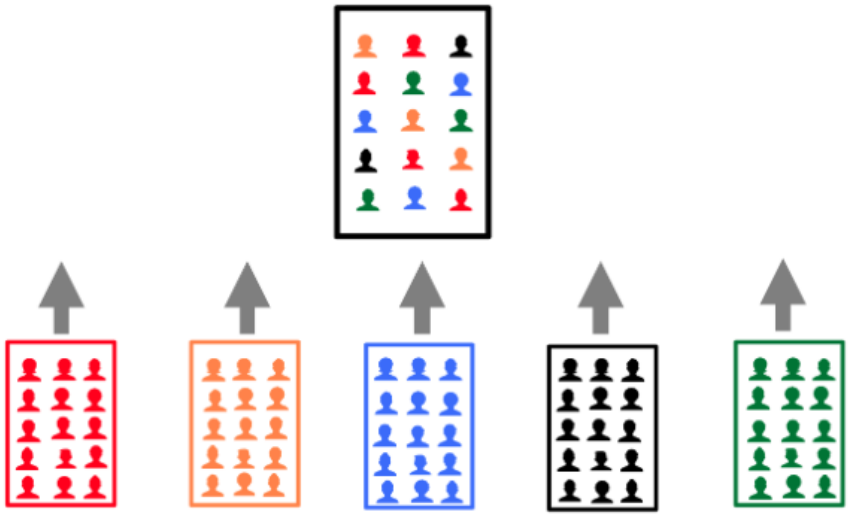
gynecologic, and urologic procedures Illinois, New York, and Texas[59]. Template matching has also been used to compare hospitals on their resource allocation of children with complex chronic conditions[62] and to compare hospital practice style and resource utilization in treating pediatric asthma patients[69]. Vincent et al. (2019) employed template matching in a more heterogeneous population, the nationwide Veterans Affairs medical, surgical, and psychiatric hospitalizations for one year[63]. It was found that hospitals differed significantly in the covariate distributions of their patients and that the limited covariate overlap between hospitals made adequate matching impossible[63]. Thus, important covariates would have to be excluded for template matching to work in this setting[61, 63].

1. A representative sample of size N is selected from the entire hospital system to serve as the template.



Entire Hospital System Population Representative Template

2. Each hospital is matched to the same template.



A similar patient at each hospital is found to match the overall template.

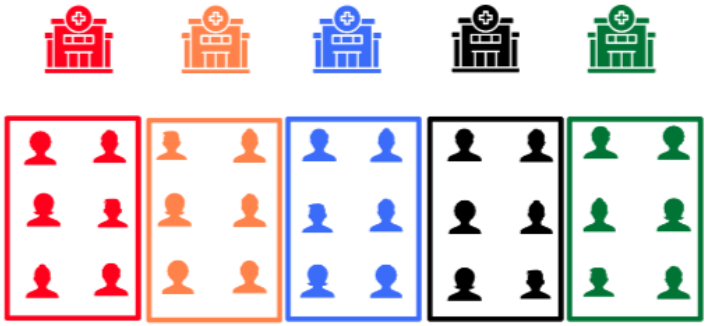
Figure 2. Diagram of Template Matching
Note: With Template Matching, each hospital is matched to a common template. All hospitals are compared on a similar set of patients. Template Matching is a form of direct standardization.

Hospital-Specific Template Matching

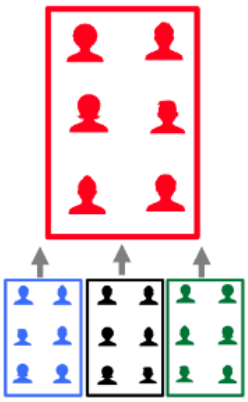
Silber et al. extended template matching to what they referred to as "hospital-specific template matching" (HS-TM)[60]. A conceptual diagram of HS-TM is presented in Figure 3. With HS-TM, each hospital has a unique template of hospitalizations that is representative of its population. Only those comparator hospitals that could be matched to each hospitalization in the template are used in the comparison. Thus, hospitals are only compared to other hospitals that treated patients like their own patients rather than all hospitals. This approach has been shown to be a feasible approach in a more heterogeneous population[64]. However, because each hospital is assessed on a sample of its population, it may be sensitive to the specific template selected or may not be representative of the hospital's total case-mix[64].

A customized comparison is in contrast with template matching[59, 63], which compares all hospitals on the same patients, regardless of how well those patients represent a given hospital's true patient population. HS-TM should not be used to rank hospitals since each hospital's performance assessment is customized to its own case-mix. When extended to the Veterans Affairs health system, this approach was feasible, and each of 122 hospitals was compared to between 6 and 64 other hospitals[64].

1. Each hospital has a template of size N that resembles its overall patient population.



2. For the hospital being evaluated (red), a similar patient is found at each of the remaining hospitals that treat similar patients (blue, black, green).



3. This process is repeated for each hospital. The hospitals are only compared to other hospitals that treated similar patients.

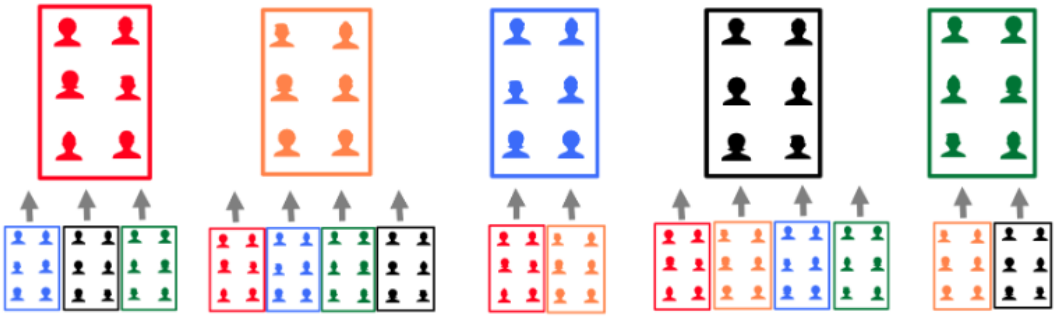


Figure 3. Diagram of Hospital-Specific Template Matching
Note: With Hospital-Specific Template Matching, each hospital has its own template and is only matched to hospitals that treated similar patients. Each hospital is then assessed by comparing the outcomes at its hospital and comparing to the outcomes of hospitals that treated similar patients. For example, the red hospital is compared on the outcomes in its template relative to the templates of the blue, black, and green hospitals.

Indirect Standardization Matching

Indirect Standardization Matching (ISM)[68] compares a given hospital's patient population to the outcomes of similar patients treated elsewhere. With ISM, each patient at the hospital undergoing the performance assessment is matched to one (or more) patient(s) from the rest of the hospital system. This approach allows a hospital to examine patient outcomes of its own patient population relative to the outcomes of similar patients treated elsewhere.

In the initial demonstration of ISM[68], Medicare patients admitted for orthopedics and common general, gynecologic, and urologic procedures in Illinois, New York, and Texas from 2004 to 2006 were included. There were 620 hospitals, and two demonstrative hospitals were used separately as the "focal" hospital (i.e., the hospital being evaluated). For every patient in the focal hospital, a ten-to-one matched comparison was found at the remaining 619 hospitals. An exact match was required for the procedure code, and optimal matching using a propensity score and Mahalanobis distance with a caliper for the propensity score and each of the risk scores was used. A caliper is the tolerable difference allowed between the treated and control groups for a match to be permitted.

For each hospital, a propensity score (i.e., the predicted probability of being a patient at the focal hospital) was derived. The propensity score model for each hospital included age, sex, emergency department admission, transfer-in status, principal procedure, and clinical risk factors. This propensity score, along with age, sex, predicted probability of 30-day mortality, predicted probability of intensive care unit admission, predicted length of stay, predicted in-hospital cost, predicted anesthesia time, emergency department admission, transfer-in status, and 21 comorbidities were used in the Mahalanobis distance. A total of 10 control individuals were then matched to each of the treated individuals using optimal matching on the Mahalanobis

distance. It is unknown how ISM performs for assessing overall hospital quality in a more diverse hospital setting.

METHODOLOGY

In this section, we present the statistical methodology and background used for the simulation study and real data application study.

Causal Inference

Rubin Causal Model

The Rubin Causal Model (RCM)[70, 71] is a framework for conducting causal inference. Causal inference is the design and analysis for evaluating the effects of a treatment or a manipulation. The "causal effect" is the comparison of different treatment conditions for the same individuals. There are two essential parts of the RCM. The first part is defining the scientific situation using "potential outcomes" to define the causal effect estimand. Causal effects are thus described by comparing the values that would be observed if the active treatment were applied to the values that would be observed if the control treatment were applied[72]. We will go into more detail about the potential outcomes framework in the next section.

The second part of the RCM is the probabilistic model for the assignment of "treatments," such as the propensity score model. The estimated propensity score[13] for individual i is the conditional probability of being assigned to the treatment given a vector of observed covariates $\boldsymbol{x}_i$:

$$\hat{P}(\boldsymbol{x}_i) = P(Z_i = 1|\boldsymbol{x}_i), \tag{1}$$

where $Z_i = 1$ for treated (and $Z_i = 0$ for control). The propensity score is most often estimated using a logistic regression model:

$$\hat{P}(\boldsymbol{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}, \tag{2}$$

for $i = 1, 2, \cdots, N$ individuals and p independent predictors. The logit of the propensity score is often used for propensity score matching, defined as:

$$\text{logit}\left(\hat{P}(X_i)\right) = \log\left(\frac{\hat{P}(\boldsymbol{x}_i)}{1-\hat{P}(\boldsymbol{x}_i)}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}. \quad (3)$$

The assignment mechanism describes why (based on the observed covariates) some individuals received the treatment, and other individuals received the control.

Potential Outcomes Framework for Causal Inference

The potential outcomes framework is a commonly used statistical framework for causal inference[73]. Only one of the potential outcomes is observed for each individual, and thus the potential outcomes framework can be thought of as a missing data problem. This is known in the causal inference literature as the "fundamental problem of causal inference"[71].

Each individual i can be potentially assigned to one of two treatment groups. Let Z_i be the binary variable indicating whether individual i is in the treated ($Z_i = 1$) or control ($Z_i = 0$) group. Additionally, we have a vector $\boldsymbol{x}_i$ of p measured covariates for each individual. Let $\boldsymbol{z} = (Z_1, Z_2, \cdots, Z_n)'$ be the $N-$vector of treatment indicators and $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \cdots, \boldsymbol{x}_n')$ be the $N \times p$ covariate matrix. Each individual has a potential outcome $Y_i(Z_i)$ under each treatment assignment. Each individual has two potential outcomes, $Y_i(1)$ and $Y_i(0)$.

To estimate the average treatment effect (ATE), we are estimating the difference between the average potential outcomes had all individuals in a population taken the treatment versus had all individuals in a population not taken the treatment. The ATE is defined as:

$$E[Y_i(1) - Y_i(0)]. \quad (4)$$

Each individual i has four quantities $\{Y_i(0), Y_i(1), Z_i, \boldsymbol{x}_i\}$, but only the outcome under the treatment they were assigned is observed (i.e., $Y_i^{obs} = Y_i(Z_i)$) and the other potential outcome is missing (i.e., $Y_i^{mis} = Y_i(1 - Z_i)$). The observed outcome is then defined as:

$$Y_i^{obs} = Y_i(1)Z_i + Y_i(0)(1 - Z_i). \quad (5)$$

The problem of causal inference is then to infer the unobserved quantities using only the observed quantities.

Assumptions of Causal Inference

One of the assumptions of causal inference is the ignorable treatment assignment mechanism[74]. This assumption states that an assignment mechanism is ignorable conditional on the covariates $\boldsymbol{x_i}$ if it does not depend on the potential outcomes. Formally,

$$P(Z_i|Y_i(0), Y_i(1), \boldsymbol{x_i}) = P(Z_i|\boldsymbol{x_i}), for\ i = 1, 2, \dots, n. \quad (6)$$

Additionally, there is a positive probability of receiving each treatment for all values of $\boldsymbol{x}$, stated formally as:

$$0 < P(Z_i = 1|\ \boldsymbol{x_i}) < 1, \forall \boldsymbol{x_i}, for\ i = 1, 2, \dots, n. \quad (7)$$

Another assumption is the Stable Unit Treatment Value Assumption (SUTVA)[1, 2, 75]. The SUTVA states that the potential outcomes for any individual do not vary with the treatments assigned to any other individual. The SUTVA is the implicit assumption under the potential outcomes framework[76]. Without SUTVA, the dimensionality of potential outcomes for each individual can be easily unmanageable even for a handful of observations. For example, suppose we have two individuals with observed outcomes and treatment assignment given as (Y_1, Z_1) and (Y_2, Z_2). If the potential outcomes of the first individual (Y_1) depended not only on Z_1 but also on Z_2, then there would four scenarios ($Z_1 \in (0,1)$ and $Z_2 \in (0,1)$). We would need four potential outcomes for Y_1 and four potential outcomes for Y_2. The number of potential outcomes needed would increase as the number of individuals increase.

Doubly Robust Approach for Assessing the Treatment Effect

Model misspecification can lead to biased estimates when using a regression model, but it is impossible to check this assumption in practice because the true relationship between the

dependent and independent variables is unknown[77]. Although we balance the included covariates within a tolerable threshold with matching, imbalance in the covariates between treatment groups may remain. To overcome some of these limitations, matching and regression can be combined in what is known as a doubly robust approach[77–81]. With propensity score matching, doubly robust estimation will yield accurate estimates of the treatment effect if *either* the propensity score model or the outcome model are correctly specified[79, 80, 82, 83]. Bias due to unmeasured confounders would be reduced if the unmeasured confounders were correlated with measured confounders included in the regression model and/or the matching algorithm[77].

We present the definition[83] of the doubly robust estimator in the context of propensity score matching because propensity score matching is one of the first matching methods used. The same ideas can then be extended to other forms of matching.

The doubly robust estimator for the ATE is:

$$\widehat{ATE} = \hat{E}\left[Y_i(1)\right] - \hat{E}\left[Y_i(0)\right]$$

$$= \frac{1}{N}\sum\left(\sum \frac{Z_i\left(Y_i - \hat{\mu}_1(\boldsymbol{x}_i)\right)}{\hat{P}(\boldsymbol{x}_i)} + \hat{\mu}_1(\boldsymbol{x}_i)\right) - \frac{1}{N}\sum\left(\frac{(1 - Z_i)\left(Y_i - \hat{\mu}_0(\boldsymbol{x}_i)\right)}{1 - \hat{P}(\boldsymbol{x}_i)} + \hat{\mu}_0(\boldsymbol{x}_i)\right), \quad (8)$$

where $\hat{P}(\boldsymbol{x}_i)$ is the estimated propensity score (through logistic regression), $\hat{\mu}_1(x)$ is the estimation of the potential outcome under the treated condition ($Y_i(1)$) using linear regression, and $\hat{\mu}_0(x)$ is the estimation of the potential outcome under the control condition ($Y_i(0)$) using linear regression. To show that the doubly robust estimator will be unbiased if either the propensity score model or the outcome model were correctly specified, consider the first part of Equation 8 without loss of generality:

$$\hat{E}[Y(1)] = \frac{1}{N}\sum\left(\frac{Z_i\left(Y_i - \hat{\mu}_1(\boldsymbol{x}_i)\right)}{\hat{P}(\boldsymbol{x}_i)} + \hat{\mu}_1(\boldsymbol{x}_i)\right). \quad (9)$$

If we assume that the specification of the outcome model $\hat{\mu}_1(x)$ is correct, then $E[Z_i\left(Y_i - \hat{\mu}_1(\boldsymbol{x}_i)\right)] = 0$. This is because multiplication of Z_i selects only the treated individuals, and by definition, the residual of $\hat{\mu}_1$ (i.e., $Y_i - \hat{\mu}_1(\boldsymbol{x}_i)$) on the treated have a mean of 0. Therefore, the reliance on the propensity score is eliminated and so it does not matter whether the propensity score model was correctly specified. Equation 9 becomes:

$$\hat{E}[Y(1)] = \frac{1}{N}\sum\left(\hat{\mu}_1(\boldsymbol{x}_i)\right), \quad (10)$$

which is $E[Y(1)]$ by assumption, and thus the estimator is unbiased. The same logic can then be extended to $\hat{E}[Y(0)]$.

Now, let us consider if instead the propensity score model was correctly specified. Equation 9 can be re-arranged as follows:

$$\hat{E}[Y(1)] = \frac{1}{N}\sum\left(\frac{Z_i\left(Y_i - \hat{\mu}_1(\boldsymbol{x}_i)\right)}{\hat{P}(\boldsymbol{x}_i)} + \hat{\mu}_1(\boldsymbol{x}_i)\right) \quad (11)$$

$$= \frac{1}{N}\sum\left(\frac{Z_iY_i}{\hat{P}(\boldsymbol{x}_i)} - \frac{Z_i\hat{\mu}_1(\boldsymbol{x}_i)}{\hat{P}(\boldsymbol{x}_i)} + \hat{\mu}_1(\boldsymbol{x}_i)\right) \quad (12)$$

$$= \frac{1}{N}\sum\left(\frac{Z_iY_i}{\hat{P}(\boldsymbol{x}_i)} - \left(\frac{Z_i}{\hat{P}(\boldsymbol{x}_i)} - 1\right)\hat{\mu}_1(\boldsymbol{x}_i)\right) \quad (13)$$

$$= \frac{1}{N}\sum\left(\frac{Z_iY_i}{\hat{P}(\boldsymbol{x}_i)} - \left(\frac{Z_i}{\hat{P}(\boldsymbol{x}_i)} - 1\right)\hat{\mu}_1(\boldsymbol{x}_i)\right). \quad (14)$$

If the propensity score model, $\hat{P}(\boldsymbol{x}_i)$, is correctly specified, then $E[Z_i - \hat{P}(\boldsymbol{x}_i)] = 0$, and therefore the reliance on the outcome model, $\hat{\mu}_1(\boldsymbol{x}_i)$, is eliminated. Equation 14 would then

reduce to the mean of the propensity score weighting estimator $\frac{Z_i Y_i}{\hat{P}(x_i)}$, which was correct by assumption. Thus, $\hat{E}[Y(1)]$ would be unbiased. The same logic can then be extended to $\hat{E}[Y(0)]$.

Mathematical Optimization

Linear programming, a form of mathematical optimization, is the problem of minimizing a linear cost function subject to linear equality and inequality constraints[84]. In a general linear programming problem, we have a vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$, called the *linear cost vector*, which is a vector of the coefficients of the objective function. We also have a *linear constraint matrix* **A** and a vector $\boldsymbol{b} = (b_1, b_2, \ldots b_m)'$ of linear constraints. Additionally, we have a vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$. The variables $x_1, \ldots, x_n$ are called *decision variables*. With linear programming, we minimize a *linear objective function* $f(x) = \boldsymbol{cx} = c_1x_1 + c_2x_2 + \cdots + c_nx_n$, subject to a set of linear equality and inequality constraints $\boldsymbol{Ax} \leq \boldsymbol{b}$. A vector **x** satisfying all constraints is called a *feasible solution*. The feasible solution that minimizes the objective function is called the *optimal solution*.

There are a variety of solvers available that can be used to solve optimization problems, including CPLEX, XPRESS, GLPK, and Symphony. However, Gurobi outperforms others, and can solve problems that others cannot and in less time[85, 86]. The Gurobi package is a commercial program, but there is a free academic license.

Cardinality Matching

With cardinality matching, there are initially T treated individuals and C control individuals and the total number of individuals in the treated and control groups is N. Each treated individual t and control individual c has a vector of observed covariates $\boldsymbol{x}$. A cardinality match is a solution to the optimization problem. The linear objective vector, $\boldsymbol{a}$, is the vector of

matched pairs between the treated and the control group. Let $a_{tc} = 1$ if treated individual t is initially matched to control individual c, and $a_{tc} = 0$ otherwise. The goal is to find $\boldsymbol{a} = (a_{11}, a_{12}, \cdots, a_{TC})$ as the solution to:

$$max \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc} \tag{15}$$

$$\text{subject to } a_{tc} \in \{0,1\},\ t = 1, 2, \cdots, T,\ \ c = 1, 2, \cdots, C$$

$$\sum_{t=1}^{T} a_{tc} \ \leq 1 \text{ for } c = 1, 2, \cdots, C,$$

$$\sum_{c=1}^{C} a_{tc} \ \leq 1 \text{ for } t = 1, 2, \cdots, T,$$

$$\mathbb{B}_k,\ k = 1, 2, \cdots, K.$$

In words, the solution to the cardinality matching problem is the largest matched sample that meets the user's supplied balance constraints. The constraint $\sum_{t=1}^{T} a_{tc} \leq 1\ for\ c = 1, 2, \cdots, C$ requires that each control individual can be used at most one time, and the constraint $\sum_{c=1}^{C} a_{tc} \leq 1\ \ for\ t = 1, 2, \cdots, T$ requires that each treated individual can be used at most one time.

As outlined in Zubizarreta, 2014, the covariate balance constraint $\mathbb{B}_k$ is a linear inequality constraint[9]:

$$\mathbb{B}_k: -b_k \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc} \leq \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc} v_{ktc} \leq b_k \sum_{t=1}^{T} \sum_{c=1}^{C} a_{tc}, \tag{16}$$

where v_{ktc} is a function of observed covariates and $b_k \geq 0$ is a given constant. We use v_{ktc} in the form $v_{ktc} = f(x_t) - f(x_c)$, for some function $f(\cdot)$. For example, let event M indicate male gender, then $f(\cdot)$ would be the indicator function for male gender defined as:

$$f(\cdot) = \mathbb{I}_{\boldsymbol{M}}(x) \coloneqq \begin{cases} 1 & if\ x \in M \\ 0 & if\ x \notin M \end{cases} \tag{17}$$

and then

$$v_{ktc} = f(x_t) - f(x_c) = \mathbb{I}_M(x_t) - \mathbb{I}_M(x_c). \quad (18)$$

Therefore, the proportion of males differs by at most b_k between the treated and control groups. If $b_k = 0$, then the matched controls would have the same number of males as in the treated group, without consideration about which specific individuals are matched. So, for example, the treated and matched control groups will have the same number of males, but a male and a female may be matched to each other. Matching exactly on a covariate in this way is known as "fine balance"[32]. Alternatively, if we let $b_k = 0.01$, for example, we limit the imbalance of the condition by at most 1%, which is referred to as "near fine balance"[87]. Or $f(\cdot)$ could be the mean age, requiring that the mean age between the treated and control groups differ by at most b_k.

The covariate balance constraint $\mathbb{B}_k$ says that the mean of v_{ktc} is within $[-b_k, b_k]$ for the matched individuals (matched individuals defined by $a_{tc} = 1$). This can be shown by rearranging the balance constraint formula in Equation 16:

$$-b_k \leq \frac{(\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc} v_{ktc})}{(\sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc})} \leq b_k. \quad (19)$$

Profile matching could also be used to identify an optimal sample that is representative of a population of interest[88], but it has not yet been used in this setting. Rather than finding a representative treated group and control group that resemble a profile of interest, the objective function could instead be changed to find the largest possible representative sample from a population. Consider the setting in which data were not uniformly sampled from a population or were sampled from a different population than the population of interest. Typically, sampling weights are used so that the resulting weighted distribution is representative of the population of

interest[1, 89, 90]. Instead, we may consider subsampling from the larger population so that the selected sample is representative of the target population.

Profile Matching

The original implementation of cardinality matching was with the R package designmatch[11]. With designmatch, the user specifies the target and tolerances. For example, the user could specify the covariate means of the treated group as the target and set the tolerance to 0.05. The treated and control groups would then be matched to the target; thus, the entire treated group would be retained. In this example, the estimand is the "ATT," or the average treatment effect on the treated[1]. Instead, the user could set the target to the means of the entire group (treated and control), and the estimand would be the "ATE" or the average treatment effect[1]. The ability to specify the target directly makes the approach very flexible and in fact, the user could instead choose an external population to serve as the target to generalize inferences to another population, a method known as profile matching[10]. Additionally, the target is not limited to means. The user could also choose to impose balance constraints on higher order moments of the covariates. The limitation of designmatch is that it only finds a one-to-one match between the treatment groups.

The MatchIt[12] R package has also added cardinality matching to its suite of matching algorithms. It has the added benefit of allowing the user to specify many-to-one matching. Additionally, it can be used to find the largest possible subset that is balanced with respect to either the treated group, control group, or overall sample. For example, if the ATT is the estimand of interest, MatchIt can be used to select the largest possible control group that is balanced with respect to the treated group and keeps the treated group intact. For the ATE, the

largest possible sample from the full dataset would be selected that is balanced with the combined treated and control group.

Profile matching could also be used to identify an optimal sample that is representative of a population of interest[10], but it has not yet been used in this setting. Rather than finding a representative treated group and control group that resemble a profile of interest, the objective function could instead be changed to find the largest possible representative sample from a population. Consider the setting in which data were not uniformly sampled from a population or were sampled from a different population than the population of interest. Typically, sampling weights are used so that the resulting weighted distribution is representative of the population of interest[1, 89, 90]. Instead, we may consider subsampling from the larger population so that the selected sample is representative of the target population.

Linear Mixed Effects Models for Multilevel Data

A simple linear regression model[91] takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{20}$$

for $i = 1, \dots, n$ individuals. This model can be extended to a multiple linear regression model which links a response to p independent predictors:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i, \text{for } i = 1, 2, \cdots, n. \tag{21}$$

One of the assumptions of standard linear regression is that the error terms of individuals are independent of one another, and thus, after accounting for the independent variables there are no relationships between the individuals[92]. However, this assumption is violated in the setting of multilevel data because individuals within a cluster (such as schools or hospitals) tend to be more alike one another than individuals between clusters. Not considering the correlation that exists between individuals within clusters can lead to underestimated standard errors because the

total variance is underestimated[93]. There are several reasons why individuals within a cluster tend to be more similar[94]. In some cases, individuals self-select their cluster membership, such as within neighborhoods or doctors' offices and thus there may be other similarities between individuals that select the same cluster. There is also potential for all members of a cluster to be affected simultaneously by cluster-level variables, such as the skill of a particular physician impacting patient outcomes. Additionally, members of the same cluster may interact and therefore influence each other.

Linear mixed effects models (also commonly referred to as multilevel models or hierarchical models) are an extension of general linear models. The data are structured in groups and coefficients can vary by groups[29]. Linear mixed effects models are widely used when data have a multilevel structure, such as patients in hospitals or students in schools[95].

There are different ways to model linear mixed effects models, including random intercept models and/or random-slope models[29]. The basic form of a random intercept model for an individual i in cluster j for is:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + \epsilon_{ij}, \quad (22)$$

where y_{ij} is the dependent variable, β_{0j} is the random intercept for cluster j (and thus varies by cluster), $\beta_1 \ldots \beta_p$ are the fixed slopes for the independent variables and ϵ_{ij} is the random error.

For a random slopes model:

$$y_{ij} = \beta_0 + \beta_{1j} x_{1ij} + \beta_{2j} x_{2ij} + \cdots + \beta_{pj} x_{pij} + \epsilon_{ij}, \quad (23)$$

and thus, the random slopes $\beta_{1j} \ldots \beta_{pj}$ vary by cluster, but the intercept β_0 is constant. In a random slopes model, the random slope is the interaction between the independent variable and

the cluster indicator[29]. A model can also have a random slope and a random intercept and thus would be in the form:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \cdots + \beta_{pj}x_{pij} + \epsilon_{ij}. \quad (24)$$

Assumptions of Linear Mixed Effects Models

With linear mixed effects models, we assume that the variance between clusters is independent of one another (i.e., the random intercept and random slopes are independent across clusters). Additionally, we assume that the cluster-level errors are independent of the individual-level errors. As with typical regression, we also assume that the residual variance is normally distributed with a mean of 0 and variance σ^2. Finally, the random intercept and slope(s) are assumed to have a multivariate normal distribution with a constant covariance matrix.

Intraclass Correlation Coefficient

The amount of correlation that exists between individuals within clusters can be estimated by using the intraclass correlation (ICC). The ICC is the proportion of the variation in the outcome variable that occurs between clusters relative to the total variation of the outcome[92]. The ICC ranges from 0, which indicates that there is no variance between clusters and the grouping contains no additional information, and 1, which indicates that there is no variance within clusters and all values within the cluster are the same[29, 92]. The ICC can be thought of as the amount of correlation for the dependent variable for two randomly selected individuals from the same cluster. Or alternatively, the amount of variation in the dependent variable can be attributed to the cluster rather than individuals. In the context of patients nested in hospitals, the cluster-level variance can be thought of as the hospital-level effect, and the ICC is the amount of variation in the outcome that can be attributed to the hospital rather than the patient. Formally,

$$ICC = \frac{\sigma_j^2}{\sigma_w^2 + \sigma_j^2}, \quad (25)$$

where σ_j^2 is the between-cluster variance and σ^2 is the within-cluster variance, and thus $\sigma_w^2 + \sigma_j^2$ is the total variance of the outcome[29].

Matching with Multilevel Data

The use of matching in the context of multilevel data requires special considerations. Thoemmes and West (2011) extended propensity score matching to clustered data in two contexts[8]. The first is when the cluster level is the central feature of the design. An example is students within schools are randomized to treatment. There may be variations in treatment implementation and interaction between students within the schools. The treatment effect of individuals (students) within the cluster (school) and across clusters is the main interest of the study. The propensity score analysis attempts to approximate a multisite randomized trial in which individuals are randomized within individual clusters.

The second scenario is when the cluster is incidental to the experimental design, and the desire is to "adjust for" the clustering. An example is randomly selected individuals who are randomized to complete a training that is delivered in a group setting. In this case, the treatment effect of the population is the focus of the study. And thus, the propensity score analysis attempts to replicate a single-level randomized experiment on individuals who are consequently clustered. However, there has been little work in examining matching designs in which the "treatment" is given to the cluster itself, and thus the cluster is the unit of analysis.

Assessing Balance After Matching

A commonly used numeric balancing diagnostics is the standardized mean difference (SMD). The SMD for a given covariate m is defined as:

$$SMD_m = \frac{|\bar{x}_{m1} - \bar{x}_{m0}|}{\sqrt{\frac{s_{m1}^2 + s_{m0}^2}{2}}}, \tag{26}$$

where $\bar{x}_{m1}$ is the mean of covariate m for the treated group, $\bar{x}_{m0}$ is the mean of covariate m for the control group, s_{m1}^2 is the variance of covariate m for the treated group, and s_{m0}^2 is the variance of covariate m for the control group. Thus, the numerator is the absolute difference in means between the treated and control groups and the denominator is the pooled standard deviation. For a binary variable, the SMD for a given covariate m is defined as:

$$SMD_m = \frac{|\hat{p}_{m1} - \hat{p}_{m0}|}{\sqrt{\frac{\hat{p}_{m1}(1 - \hat{p}_{m1}) + \hat{p}_{m0}(1 - \hat{p}_{m0})}{2}}}, \tag{27}$$

where $\hat{p}_{m1}$ is the proportion of covariate m for the treated group and $\hat{p}_{m0}$ is the proportion of covariate m for the control group. Thus, the numerator is the absolute difference in proportions between the treated and control groups and the denominator is the pooled standard deviation. Higher values of SMD indicate a greater imbalance across covariates. There are different recommendations for thresholds for SMD that are considered "balanced," including 0.1[24, 96] or 0.25 for regression to be trustworthy[1, 97].

For many-to-one matching, a weighted SMD can be used to assess balance[98]. The weighted mean is defined as:

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}, \tag{28}$$

while the weighted sample variance is defined as:

$$s_w^2 = \frac{\sum w_i}{(\sum w_i)^2 - \sum w_i^2} \sum w_i \, (x_i - \bar{x}_w)^2. \tag{29}$$

Each treated individual is given a weight of 1 and each matched control is given a weight w_i equal to the reciprocal of the number of matched controls. For one-to-one matching, the

weights are 1 for both the treated and control individuals. For 2-to-1 matching $w_i = 1/2$ for the matched controls and $w_i = 1$ for the treated individuals and for ten-to-one matching, $w_i = 1/10$ for the matched controls and $w_i = 1$ for the treated individuals. The weighted means and weighted variances are computed separately for the treated group and matched control group, and then used in Equation 26 to compute the weighted SMD for continuous variables.

SIMULATION STUDY

The objective was to examine the impact of increasing the number of matched controls on the estimation of the treatment effect for studies conducting profile matching with multilevel data. We considered 1 through 10 matched controls and the largest possible representative subset[99]. Additionally, we compared using a linear model to a linear mixed effects model after matching.

Methods

Data Generation

We simulated a population of 100,000 individuals nested in clusters, comprised of 1,000 "treated" and 99,000 "control" individuals. We considered the scenario in which the entire treated group came from a single cluster, such as a school or hospital. The control group was comprised of individuals from 99 other clusters. We randomly assigned individuals to the 99 clusters with equal probability, so that the clusters would be similar in size, but not exactly equal size. For each cluster, we generated a random cluster-level intercept γ_j from a normal distribution, such that the intraclass correlation (ICC) was 0.03, 0.05, or 0.10, based on ICCs reported in other studies[100–103]. In the context of a model with patients nested in hospitals, the ICC is the proportion of the variation in the patient outcome that can be explained by the hospital rather than the patient characteristics.

For each individual, we generated 10 normally distributed variables to represent the measured covariates. Variables from the treated and control groups were generated from different distributions to control the amount of covariate overlap between the groups. The covariates for the control group came from the standard normal distribution. The covariates for the treated group came from a normal distribution with a variance of 1 and with the mean shifted,

such that the mean of the SMD of the covariates between the treated and control groups was the desired value.

For our simulated data, since each covariate in the control group has a mean of 0 and both groups have a variance of 1, the SMD is the absolute difference of the means of the covariates between the treated and control groups. We varied the means for the covariates in the treated group so that the mean SMD was either 0.3, 0.4, or 0.5, depending on the scenario.

We also simulated one additional covariate from the standard normal distribution for both the treated and control groups to represent an unmeasured covariate. The data were simulated so that the measured covariates explained approximately 80% of the variance in the outcome. There was no treatment effect, and therefore, any differences in the outcomes between the treatment groups were due to differences in the covariate distributions.

Without loss of generality, the covariate overlap of a single covariate x_1 is depicted in Figure 4 when the SMD is 0.1, 0.3, and 0.5. On the right, we display the corresponding overlap in the propensity score using all 10 measured covariates. When the SMD for a given covariate is 0.1, which is typically considered "balanced," we see a fair overlap in the propensity score. As the SMD increases, the distribution of the propensity scores starts to separate. There is still overlap between the two groups when the SMD is 0.5, but the overlap is limited. This implies that finding sufficient matches will be more difficult.

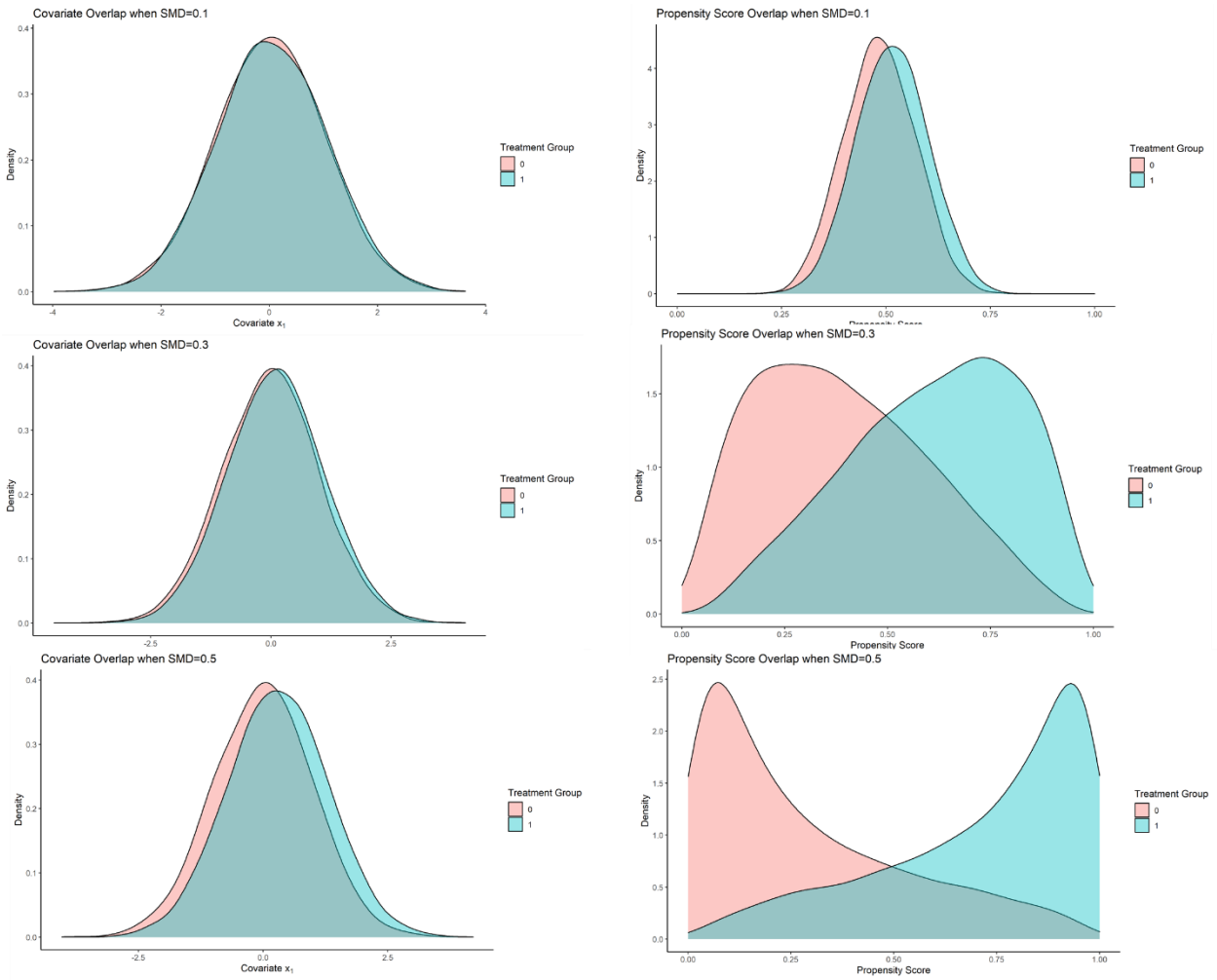


Figure 4. Covariate Overlap and Propensity Score Overlap
Note: On the left, we display the covariate overlap for the treated group (blue) and the control group (pink) for a single covariate. The standardized mean difference (SMD) between the treated and control groups is 0.1 (row 1), 0.3 (row 2), and 0.5 (row 3). On the right, we show the corresponding propensity score overlap when each of the 10 covariates has the given SMD.

The outcome for individual i in cluster j was:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_{11} x_{11ij} + \gamma_j + \epsilon_{ij}, \quad (30)$$

for $i = 1, 2, \cdots, 100{,}000$, and $j = 1, 2, \cdots, 100$, where $\gamma_j \overset{iid}{\sim} N(0, \sigma_j^2)$ and $\epsilon_{ij} \overset{iid}{\sim} N(0,1)$. For our simulation, we assumed that all covariates have equal contributions to the outcome, so we arbitrarily assigned the model intercept $\beta_0 = -1$ and each of the beta coefficients β_1 through β_{11} to 1 for all clusters, and thus we used a random-intercepts model. We varied the cluster-level variance σ_j^2 so that the ICC was either 0.03, 0.05, or 0.10.

The outcome is the sum of 11 independent random variables, the cluster-level variance, and the residual error variance for the simulated data. The total variance of the model is:

$$\sigma_{total}^2 = 12 + \sigma_j^2. \quad (31)$$

Thus,

$$ICC = \frac{\sigma_j^2}{12 + \sigma_j^2} \quad (32)$$

and we set the ICC to the desired level (either 0.03, 0.05, or 0.10) and solved for σ_j^2. The corresponding variance for each ICC level was 0.3711, 0.6316, and 1.3333, respectively. We then simulated γ_j from a normal distribution with a mean of 0 and a variance of σ_j^2. Overall, we conducted nine scenarios: three levels of covariate overlap in the measured covariates and three levels of ICC. Each scenario was run 1,000 times.

Matching Approach

For each simulated dataset, we conducted 1-to-1 matching, then 2-to-1 matching, etc., up to 10-to-1 matching. Additionally, we found the largest possible subset of the control group that was representative of the treated group. We included the 10 measured covariates in the matching procedure (i.e., the "unmeasured" covariate was not included in the matching

procedure). We used cardinality matching with tolerances set so that the maximum mean SMD between the treated and control groups was 0.05 for each covariate, thus requiring that each measured covariate was balanced between the treatment groups, with SMD<0.1. As a sensitivity analysis, we also simulated the data such that the tolerable differences between the treated and control group were 0.10 for each covariate. Cardinality matching was conducted using the ProfileMatchit package in R and the Gurobi optimization software[104]. We set the target to be the means of the treated group, and thus we were estimating the average treatment effect of the treated (ATT).

Assessing Covariate Balance

After matching, we computed the balance of the matched sample by determining the SMD between the treated group and matched control group for each measured covariate. We then calculated the mean of the SMD to get an overall measure of balance across the 10 covariates. For the many-to-one matched samples, we used the weighted SMD[98].

Estimating the Treatment Effect

After a matched sample was constructed, we estimated the treatment effect in two ways. First, we estimated the difference between the outcome Y in the treated and matched control groups using a linear regression model with the treatment group as the only covariate. Next, we used a linear mixed effects model in which we included the cluster as a random intercept to control for the clustering of the control group. As an additional sensitivity analysis, we also considered a doubly robust approach, wherein the models were adjusted for the 10 measured covariates in addition to the treatment indicator.

Bias, Precision, Accuracy, and Coverage

Separately for the linear and linear mixed effects models, we computed the bias of the treatment effect, defined as the difference between the estimated treatment effect and the true treatment effect (which was simulated to be 0, and thus there was no treatment effect). We reported the mean bias and 95% confidence intervals across the 1,000 simulations for each scenario. To measure the precision of the estimate, we calculated the variance of the estimated treatment effect across the 1,000 simulations. We also calculated the mean squared error (MSE) of the estimated treatment effect across the 1,000 simulations, which is a measure of the accuracy of the estimator. The MSE is the sum of the variance of the estimator and the square of the bias of the estimator[105].

Additionally, we determined the coverage of the estimator, which is the proportion of the 1,000 simulations that contained the true treatment effect in the 95% confidence interval for the estimate. Our objective in determining the optimal number of matched controls is to identify the number of matched controls that resulted in minimal bias, minimal sample variance, minimal MSE, and maximal coverage. However, trade-offs must be made between these, so we reported the number of matched controls for each scenario that minimized the MSE, minimized the measured absolute bias, minimized the sampling variance, and maximized the coverage.

Results

The covariates of the unmatched simulated data were imbalanced between the treated and control groups by design. The standardized mean difference for the unmatched data was 0.3, 0.4, or 0.5. After matching one control, the treated and control groups were balanced (as defined by having a mean SMD<0.1). The balance remained as the number of matched controls increased. Figure 5 displays the SMD of the measured covariates of the simulated data. The initial amount

of imbalance (i.e., the initial SMD) between the treated and control groups had little impact on the overall balance after matching (Figure 5). The tolerances were set at .05 standard deviations for each covariate, and thus the maximum SMD between the treatment group is 0.05. For each of the 1,000 simulations and 9 scenarios, the range of balance (SMD) in the resulting matched samples ranged between 0.032 and 0.050, while the overall mean SMD across the 1,000 simulations ranged from 0.047 to 0.050 depending on the scenario (Appendix Table A1).

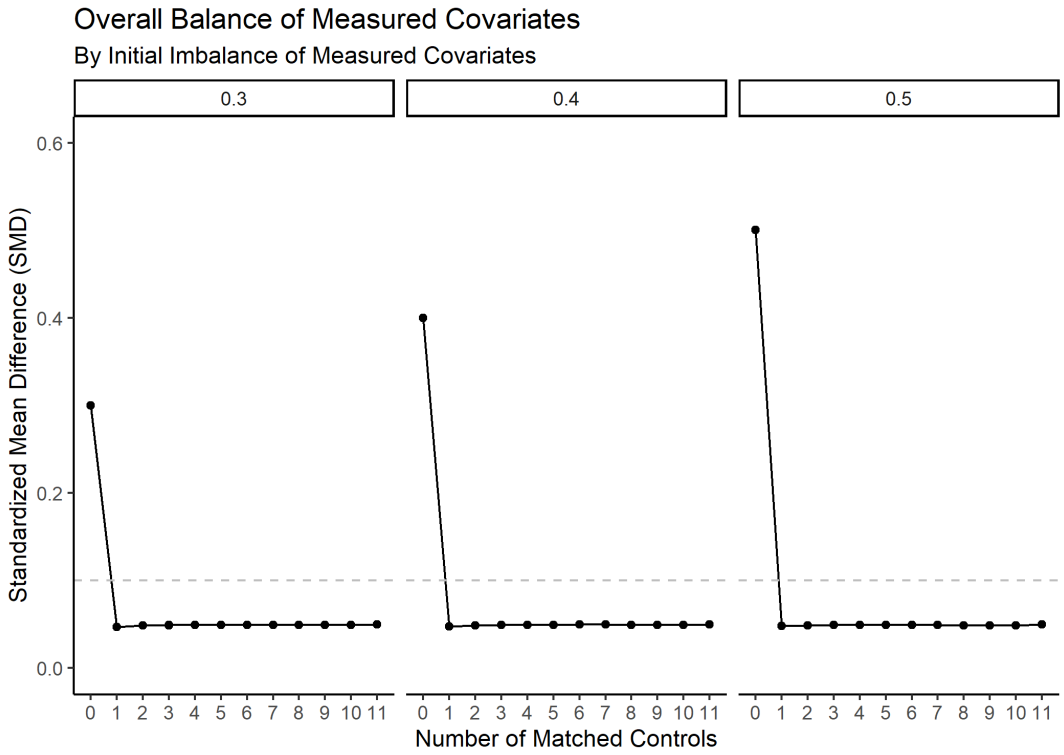


Figure 5. Standardized Mean Difference of the Overall Balance of Measured Covariates
Note: The initial imbalance of the measured covariates is provided in the inset. The unmatched data has a standardized mean difference that corresponds to the initial imbalance of 0.3, 0.4, or 0.5. The number of matched controls is depicted on the x-axis.

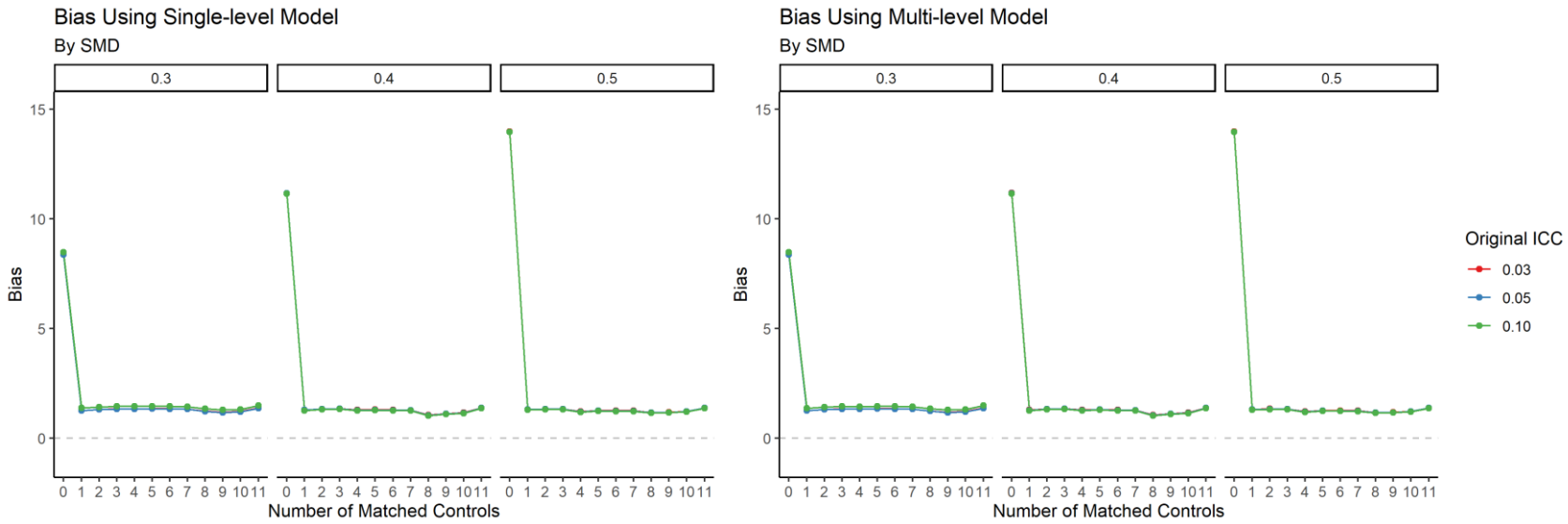


Figure 6. Bias of the Treatment Estimate

Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the bias of the treatment estimate using a single-level linear model and on the right is the bias of the treatment effect using a linear mixed effects model. The x-axis is the number of matched controls (where 11 indicates the largest possible subset), and the y-axis is the bias of the treatment estimate. The inset provides the initial simulated imbalance of the measured covariates. The colored lines denote the initial intraclass correlation.

Figure 6 displays the bias of the treatment effect for each additional matched control. Because we were able to balance the treated and control groups, the initial imbalance of the covariates did not impact the absolute bias of the treatment effect estimate. The estimated bias for each scenario for the unmatched, one through ten matched controls, and the largest subset is provided in Appendix Table A2. The linear mixed effect model resulted in estimates that were less biased than when used a linear model when matching one through ten controls. However, when using the largest possible subset, the bias was the same for a linear model and a linear mixed effects model. Indeed, the bias was worse when using a linear mixed effects model with the largest possible subset than when using a linear mixed effect model with one-to-one matching and when using many-to-one matching.

After matching to one control, the bias decreased for all scenarios by a mean of 87% (range 72% to 93%) when using a linear mixed effects model relative to the unmatched sample. Matching to an additional matched control increased the bias by a mean of 3% (range 0% to 6%). Using the largest subset resulted in an increased bias of 7% (range 5% to 10%) compared to using one match. Overall, the level of bias was consistent across the number of matched controls, with the smallest observed bias occurring at eight or nine matched controls, depending on the scenario. However, the relative reduction in bias using eight or nine controls was minimal relative to matching one control.

Figure 7 displays the coverage of the confidence interval for the treatment estimate across the 1,000 simulations. Using a linear model resulted in coverage of at most 31% across all scenarios. However, the use of a linear mixed effects model yielded adequate coverage across scenarios. After matching to one control, the coverage improved by a mean of 87 percentage points (range: 72 to 93 percentage points) across the simulations using a mixed effects model

(Appendix Table A3). Additional matches decreased the coverage by a mean of 3 percentage points for two controls up to a mean of 5 percentage points for the largest subset.

The MSE is provided in Appendix Table A4, and the variances are in Appendix Table A5. Matching to one control decreased the MSE by an average of 93% (range 85% to 98%). Increasing the number of matched controls had little impact on the MSE. The variance was reduced by a mean of 80% (range 58% to 93%) after matching to one control, and then changed marginally for each additional matched control.

We conducted two sensitivity analyses. First, we estimated the treatment effect using a doubly robust approach. With a doubly robust approach, regression adjustment is used to remove the small remaining residual covariate imbalance between treatment groups[1]. The number of matched controls had little effect on the bias, coverage, MSE, or variance of the treatment effect estimate when using a model that adjusted for covariates. In our simulations, the tolerable imbalance between treatment groups was small (SMD$\leq$ 0.05) (Appendix Table A1); however, there was a slight bias in the treatment effect estimate after matching. The use of a doubly robust approach nearly eliminated the remaining bias (Appendix Table A6). Figure 8 presents the bias of the treatment effect estimate using a model that adjusted for the measured covariates, and Figure 9 presents the coverage of the treatment effect estimate confidence interval for a model that adjusted for the measured covariates. The coverage of the models adjusted for covariates is presented in Appendix Table A7, while the MSE is presented in Appendix Table A8 and the variance in Appendix Table A9.

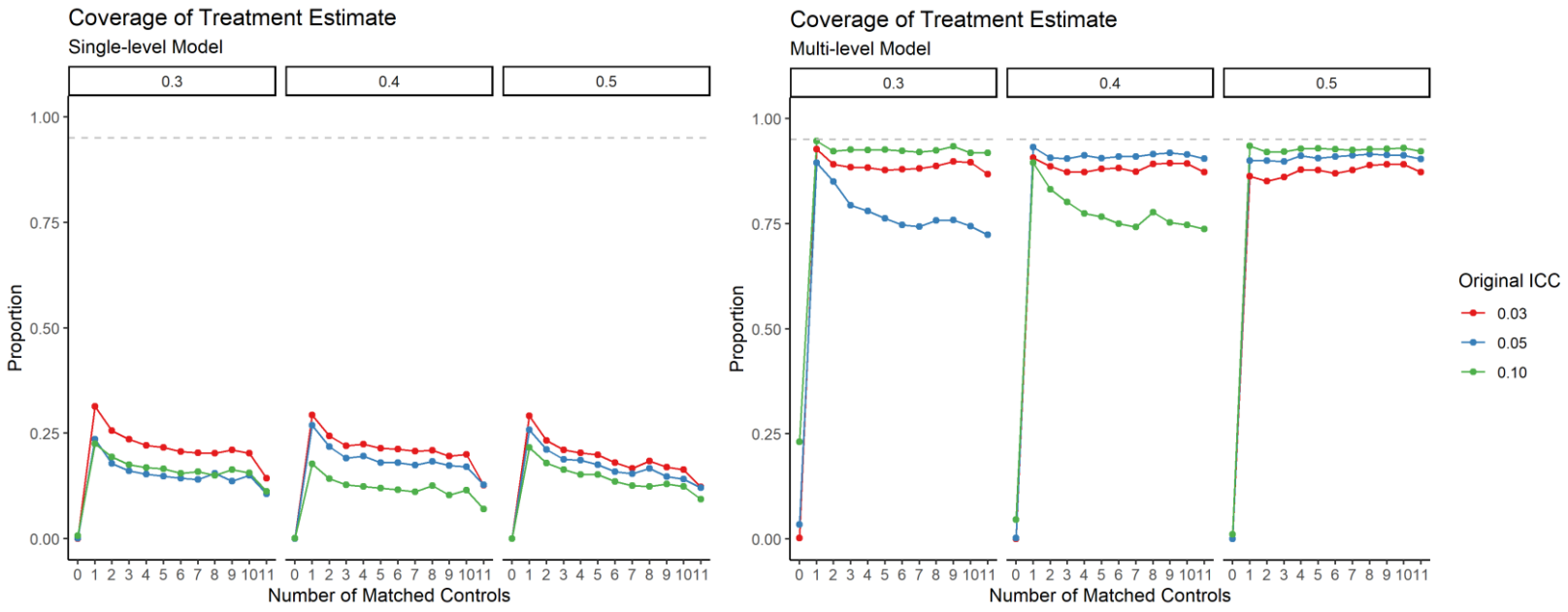


Figure 7. Coverage of the Treatment Estimate 95% Confidence Interval

Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the coverage of the treatment estimate 95% confidence interval using a single-level linear model and on the right is the coverage of the treatment effect 95% confidence interval using a linear mixed effects model. The x-axis is the number of matched controls (where 11 indicates the largest subset). The y-axis is the proportion of iterations in which the 95% confidence interval of the treatment estimate contained the true treatment effect. The inset provides the initial simulated imbalance of the covariates. The colored lines denote the intraclass correlation of the cluster.

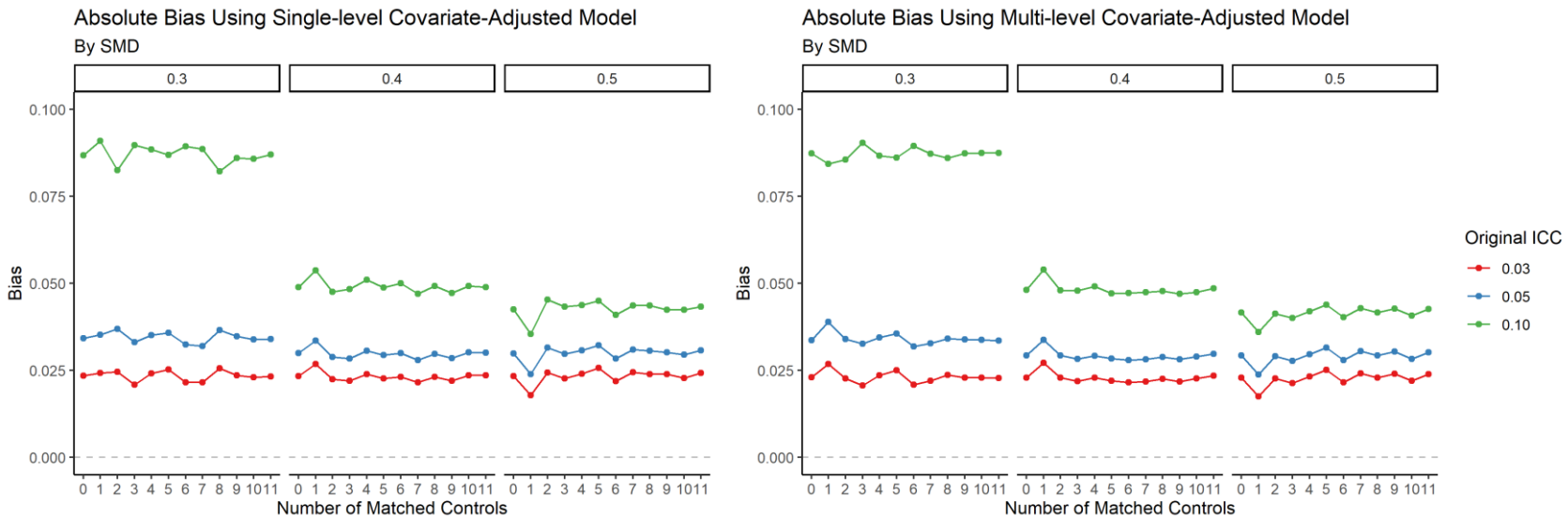


Figure 8. Bias of the Treatment Estimate Using a Covariate-Adjusted Model

Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the bias of the treatment estimate using a single-level linear model adjusted for the measured covariates and on the right is the bias of the treatment effect using a linear mixed effects models adjusted for the measured covariates. The x-axis is the number of matched controls (where 11 indicates the largest possible subset), and the y-axis is the bias of the treatment estimate. The inset provides the initial simulated imbalance of the measured covariates. The colored lines denote the initial intraclass correlation.

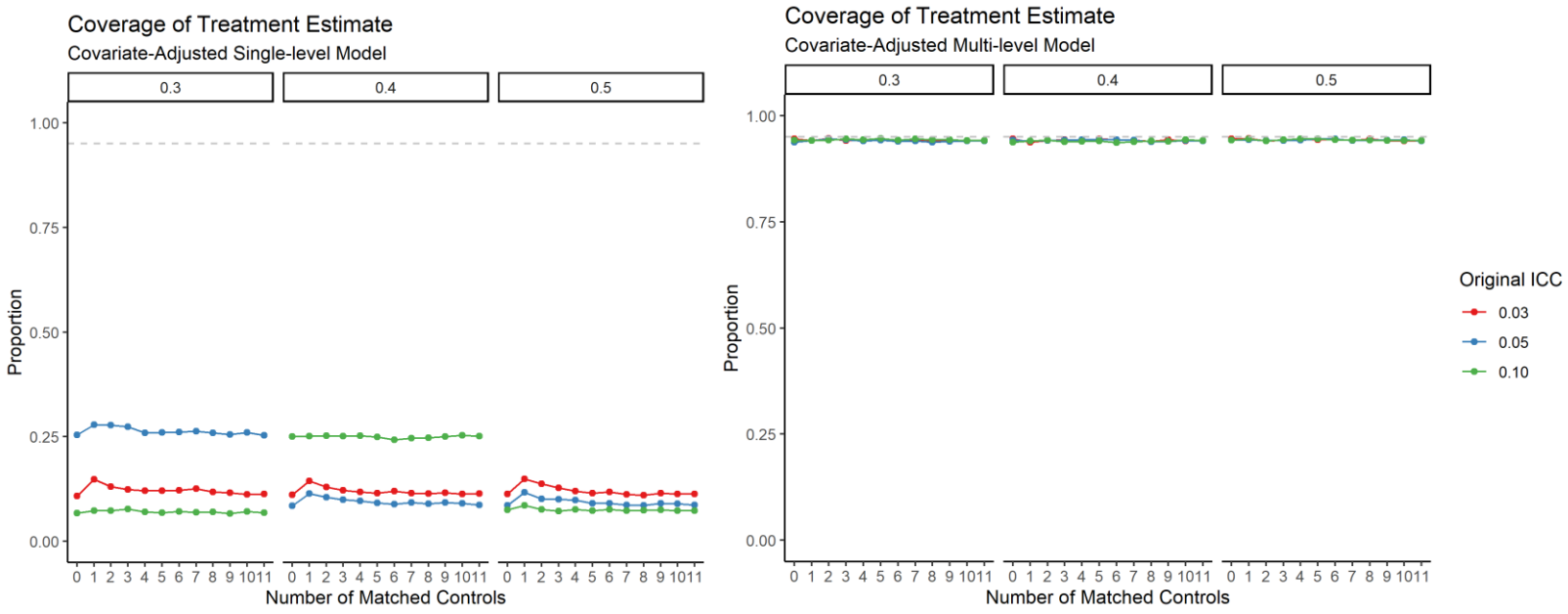


Figure 9. Coverage of the Treatment Estimate 95% Confidence Interval Using a Covariate-Adjusted Model

Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the coverage of the treatment estimate 95% confidence interval using a single-level linear model and on the right is the coverage of the treatment effect 95% confidence interval using a linear mixed effects model. The x-axis is the number of matched controls (where 11 indicates the largest subset). The y-axis is the proportion of iterations in which the 95% confidence interval of the treatment estimate contained the true treatment effect. The inset provides the initial simulated imbalance of the covariates. The colored lines denote the intraclass correlation of the cluster.

Next, we re-ran the simulations with the tolerances set at 0.10 standard deviations rather than 0.05 standard deviations. Using this threshold keeps the covariates balanced according to published guidelines[15, 28]. We found that the mathematical optimization process finds an acceptable cardinality match that resulted in a SMD very close to the tolerance level specified (Figure 10). When the threshold was set at 0.05 standard deviations, the resulting matched sample attained a balance between 0.047 and 0.050. When the threshold was set at 0.10 standard deviations, the attained balance was between 0.094 and 0.100. Matches are easier to obtain with a higher tolerance, and indeed it may not be possible to find matches with a lower tolerance. However, we found that using the higher tolerance resulted in biased estimates (Figure 11) that did not attain adequate coverage (Figure 12), even when using a mixed effect linear model. Adjusting for covariates reduced the bias (Figure 13) and increased the coverage (Figure 14), but the estimates were still biased and had coverage less than 0.95.

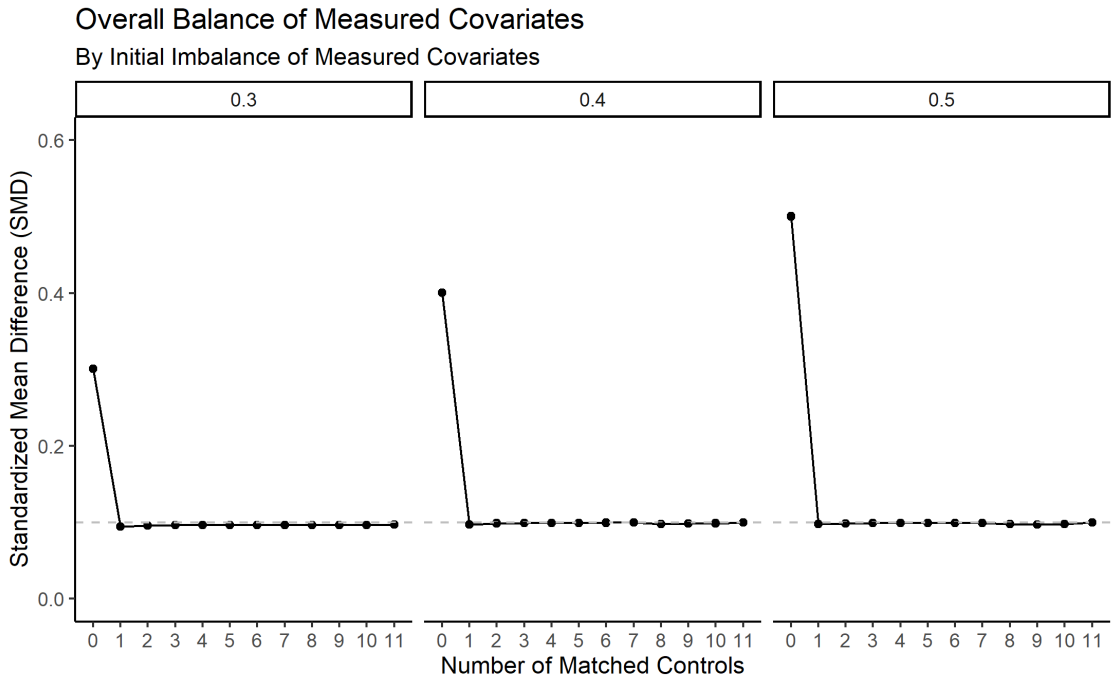


Figure 10. Standardized Mean Difference of the Overall Balance of Measured Covariates with Tolerances Set at 0.10 Standard Deviations

Note: The initial imbalance of the measured covariates is provided in the inset. The unmatched data has a standardized mean difference that corresponds to the initial imbalance of 0.3, 0.4, or 0.5. The number of matched controls is depicted on the x-axis.

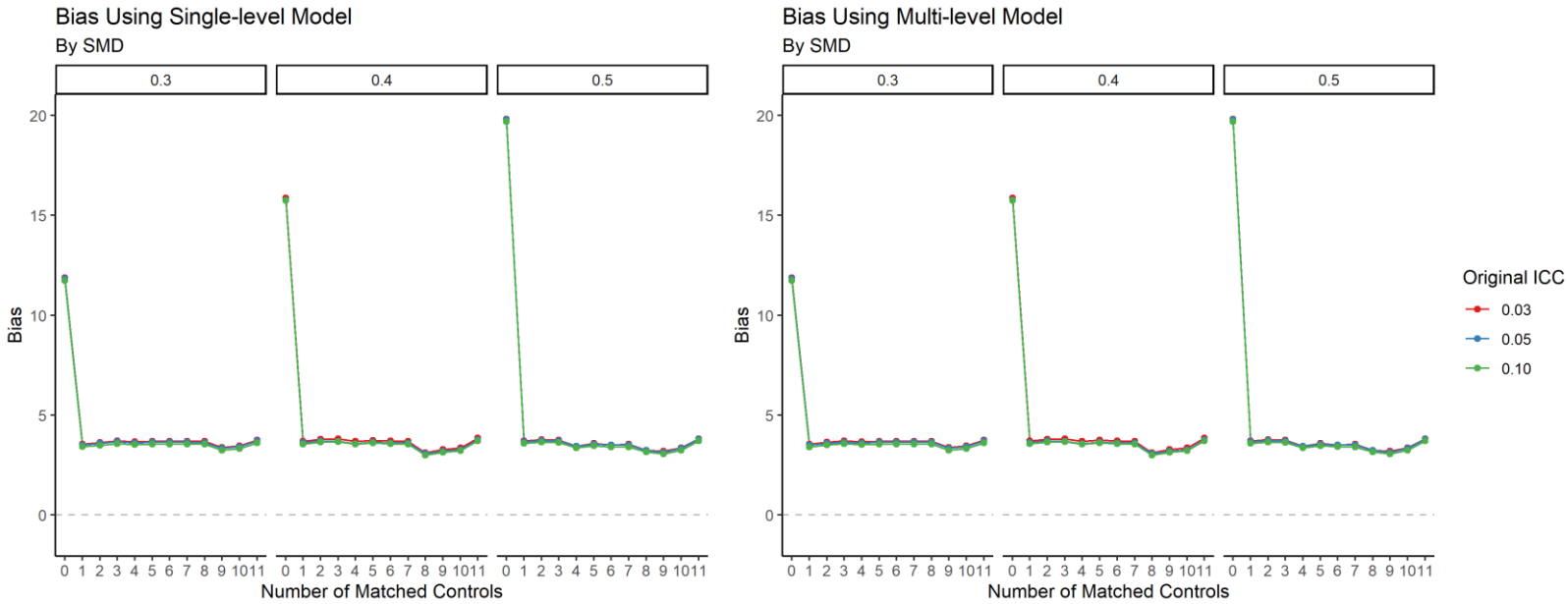


Figure 11. Bias of Treatment Effect Estimate with Tolerances Set at 0.10 Standard Deviations

Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the bias of the treatment estimate using a single-level linear model adjusted for the measured covariates and on the right is the bias of the treatment effect using a linear mixed effects models adjusted for the measured covariates. The x-axis is the number of matched controls (where 11 indicates the largest possible subset), and the y-axis is the bias of the treatment estimate. The inset provides the initial simulated imbalance of the measured covariates. The colored lines denote the initial intraclass correlation.

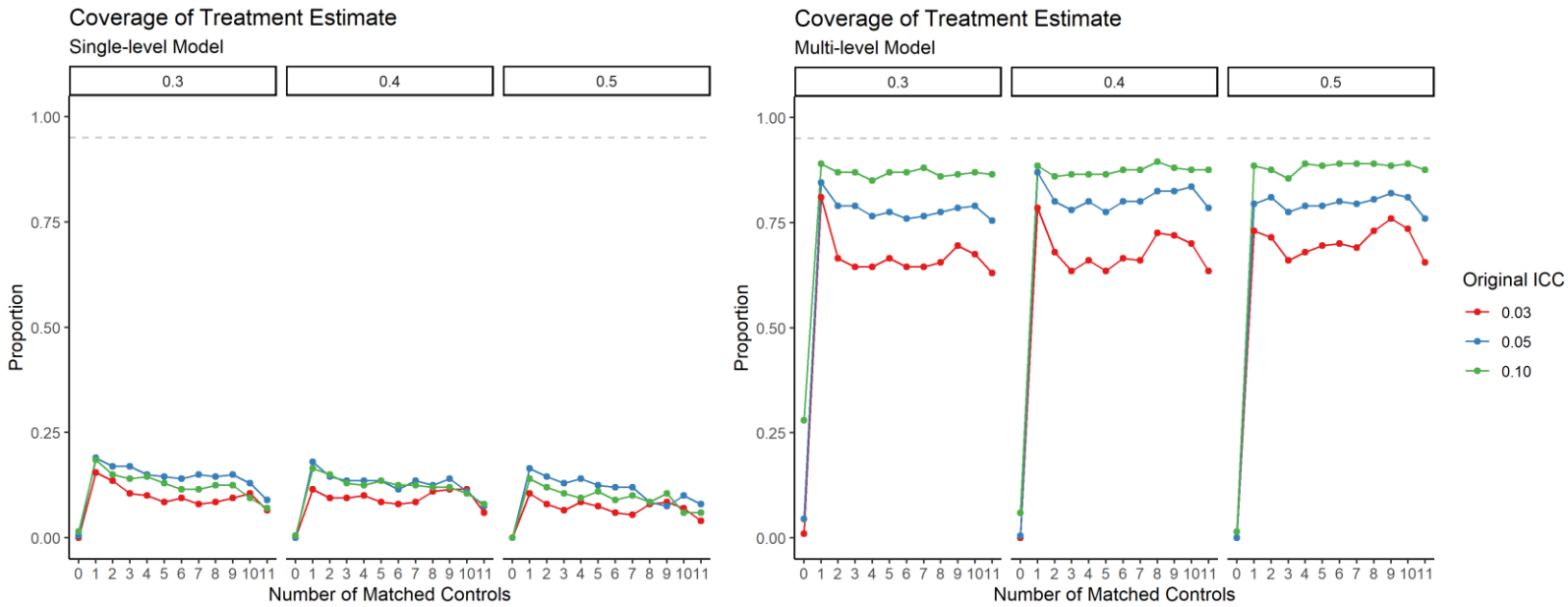


Figure 12. Coverage of the Treatment Estimate 95% Confidence Interval with Tolerances Set at 0.10 Standard Deviations
Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the coverage of the treatment estimate 95% confidence interval using a single-level linear model and on the right is the coverage of the treatment effect 95% confidence interval using a linear mixed effects model. The x-axis is the number of matched controls (where 11 indicates the largest subset). The y-axis is the proportion of iterations in which the 95% confidence interval of the treatment estimate contained the true treatment effect. The inset provides the initial simulated imbalance of the covariates. The colored lines denote the intraclass correlation of the cluster.

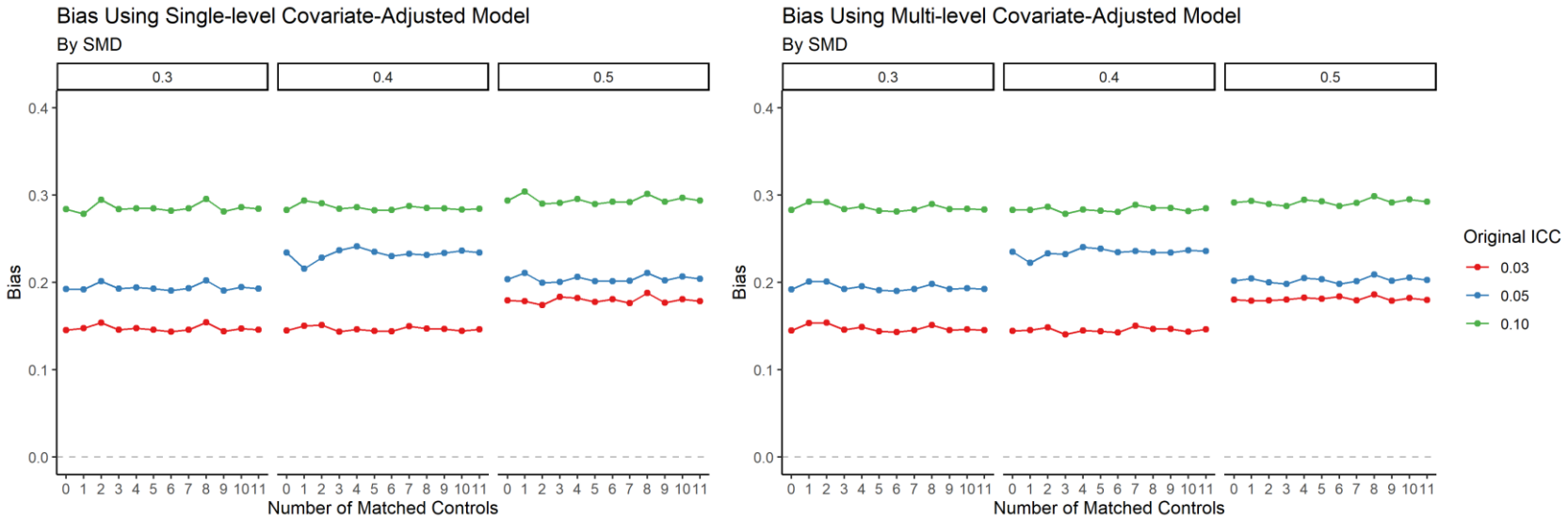


Figure 13. Bias of Treatment Effect Estimate with Tolerances Set at 0.10 Standard Deviations Using a Covariate-Adjusted Model
Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the bias of the treatment estimate using a single-level linear model adjusted for the measured covariates and on the right is the bias of the treatment effect using a linear mixed effects models adjusted for the measured covariates. The x-axis is the number of matched controls (where 11 indicates the largest possible subset), and the y-axis is the bias of the treatment estimate. The inset provides the initial simulated imbalance of the measured covariates. The colored lines denote the initial intraclass correlation.

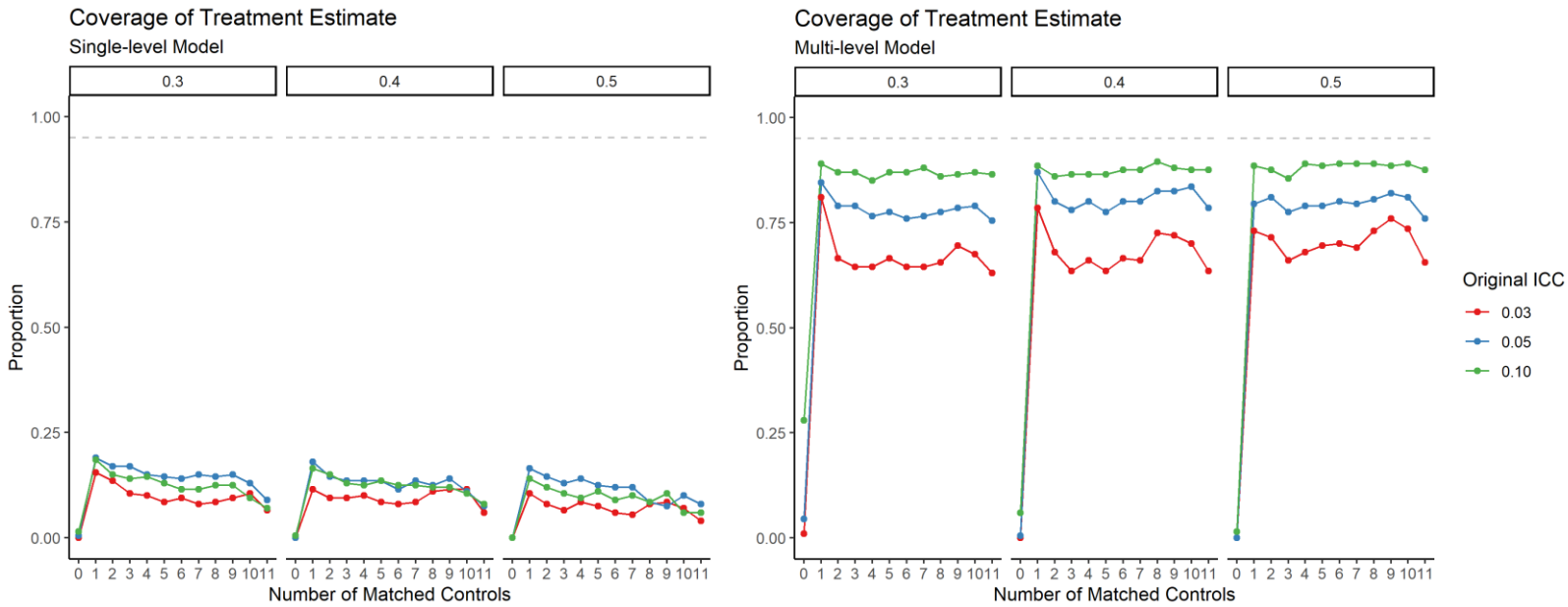


Figure 14. Coverage of the Treatment Estimate 95% Confidence Interval with Tolerances Set at 0.10 Standard Deviations Using a Covariate-Adjusted Model

Note: ICC=intraclass correlation. SMD=standardized mean difference. On the left is the coverage of the treatment estimate 95% confidence interval using a single-level linear model and on the right is the coverage of the treatment effect 95% confidence interval using a linear mixed effects model. The x-axis is the number of matched controls (where 11 indicates the largest subset). The y-axis is the proportion of iterations in which the 95% confidence interval of the treatment estimate contained the true treatment effect. The inset provides the initial simulated imbalance of the covariates. The colored lines denote the intraclass correlation of the cluster.

INDIRECT STANDARDIZATION PROFILE MATCHING FOR HOSPITAL QUALITY ASSESSMENT USING REAL DATA

The objective was to extend profile matching to perform indirect standardization matching using a newly proposed cardinality matching algorithm for the setting of hospital quality assessment. We proposed a hybrid method that combined profile matching[10] and indirect standardization matching[68], which we refer to as indirect standardization profile matching (IS-PM). Profile matching was first introduced as an approach to generalize causal inferences of a new target population or personalize causal inferences for an individual[10]. It has not yet been used in the setting of hospital performance assessment, but the methodology has a natural extension to this setting. Indirect standardization matching[68] was first introduced to conduct hospital performance assessments to overcome the limitations of standard regression-based approaches. It was proposed in a limited set of hospitals that performed general, urologic, or gynecologic surgical procedures in Medicare patients in three states. In the original implementation, 10-to-1 Mahalanobis distance matching was used. In this chapter, we used profile matching, an extension of cardinality matching[9, 26, 106], to balance the covariates in the matched samples directly.

Methods

Data Source

We used data from the eICU Collaborative Research Database[107, 108]. The eICU is a multi-center intensive care unit database for over 200,000 admissions to intensive care units monitored by eICU Programs across the United States. Highly granular patient data is included, including vital signs and laboratory measurements, the severity of illness measures, and

diagnoses. The database is de-identified and publicly available after approval. Table 1 provides a summary of the data tables available in the eICU database that were used.

Table 1. Summary of Selected Data Tables Available in the eICU Database

Concept	eICU Data Table	Description
Demographics	patient	Contains patient demographics and admission and discharge details for hospital and intensive care unit stays.
Hospital	hospital	Contains details of hospitals covered by the eICU telehealth program.
Care Plan	carePlanGeneral	Documentation relating to care planning, continuously updated over a patient stay.
APACHE score	apacheApsVar	Contains the variables used to calculate the acute physiology score III for patients.
APACHE score	apachePredVar	Provides variables underlying the APACHE predictions. APACHE consists of a group of equations used for predicting outcomes in critically ill patients. APACHE is based on the acute physiology score (which uses 12 physiologic values), age, and chronic health status within one of 56 disease groups.
APACHE score	apachePatientResult	Provides predictions made by the APACHE score (versions IV and IVa), including the probability of mortality, length of stay, and ventilation days.

Note: APACHE=Acute Physiology and Chronic Health Evaluation. APS=Acute Physiology Score.

Acute Physiology and Chronic Health Evaluation (APACHE) IV

The acute physiology and chronic health evaluation system (APACHE) is a widely known predictive tool for in-hospital mortality and length of stay for patients in critical care[109]. APACHE scores are generated using demographic factors, physiologic measures, and diagnoses from the first 24 hours of a patient's intensive care unit stay. The current version is APACHE IV, which was developed in 2006 as a recalibration and improvement over the APACHE III score model[109].

The APACHE was developed using data from approximately 110,000 admissions from 104 intensive care units. The model had an area under the receiver operating curve of 0.88 for

the prediction of hospital mortality using the validation population. It had a Hosmer-Lemeshow p-value of 0.08, where p>0.05 indicates good calibration. The Hosmer-Lemeshow test is a common method of assessing calibration and runs a chi-square test calculation to test whether the model matched the results expected from perfect calibration[110]. For a well-calibration model, the mortality probabilities should be consistent with the underlying mortality probability distribution. That is, in a well-calibrated model, approximately 50% of patients who had a predicted probability of mortality of 50% should have observed mortality.

APACHE is composed of three parts: acute physiology, age, and chronic conditions. The acute physiology score ranges from 0 to 252, the chronic conditions score ranges from 0 to 23, and the age score ranges from 0 to 24. The APACHE score is the sum of these three scores, and thus the permissible range is from 0 to 299.

Acute Physiology Score

The Acute Physiology Score (APS) III is an established method for summarizing patient severity of illness on admission to the intensive care unit. The APS variables include neurological abnormalities based on Glasgow Coma Scale[111, 112], pulse rate, mean blood pressure, temperature, respiratory rate, PaO_2/FiO_2 ratio (or $P(A-a)O_2$ for intubated patients with $FiO_2 \geq 0.5$), hematocrit, white blood cell count, creatinine, urine output, blood urea nitrogen, sodium, albumin, bilirubin, glucose, and acid-base abnormalities[109]. Each of the individual variables is available in the eICU data table "apacheApsVar." The worst recorded value *(*i.e., the value that has the highest deviation from normal) from the first 24 hours of the intensive care unit stay is recorded in raw form, so the variables need to be scored according to the APACHE IV point system. The APS is the sum of the points. Missing APS variables are assumed to be normal and thus receive an APS value of 0. The points assigned to each element

of the APS are provided in Appendix Table A10. The Glasgow coma scoring is provided in Appendix Table A11, and the APACHE scoring for the Glasgow Coma Score is in Appendix Table A12.

Chronic Conditions Score

There are seven chronic conditions included in the APACHE with corresponding points. These include AIDS (23 points), hepatic failure (16 points), lymphoma (13 points), metastatic cancer (11 points), immunosuppression (10 points), leukemia or myeloma (10 points), and cirrhosis (4 points)[113]. Comorbid conditions are excluded for elective surgery patients[113].

Age Score

The APACHE assigns weights to the age of the patient as follows[113] ≤44 (0 points), 45-59 (5 points), 60-64 (11 points), 65-69 (13 points), 70-74 (16 points), 75-84 (17 points), and ≥85 (24 points).

Other APACHE variables

The APACHE also incorporates intensive care unit admission source (floor, emergency department, operating/recovery room, stepdown unit, direct admission, other intensive care unit, another hospital, another admission source), length of stay before intensive care unit admission, emergency surgery (Y/N), thrombolytic therapy for patients with acute myocardial infarction (Y/N), mechanical ventilation (Y/N), and intensive care unit admission diagnoses.

The APACHE admission diagnoses were categorized to match the Australian and New Zealand Intensive Care Society Adult Patient Database. Admission diagnoses are either “post-operative” or “non-operative.” The APACHE III-J diagnoses categories include cardiovascular, respiratory, gastrointestinal, neurological, sepsis, trauma, metabolic, hematological/endocrine, renal/genitourinary, musculoskeletal/skin, and gynecological[114].

Exclusion

APACHE hospital mortality prediction is not conducted for patients less than 16 years old, burn patients, in-hospital intensive care unit readmissions, transplant patients (except hepatic and renal transplants), patients with a length of stay >365 days, patients with a length of stay <4 hours, or if there is no diagnosis within the first day of the intensive care unit. Thus, we excluded all patients for which an APACHE prediction was not made for this study, consistent with other researchers[115, 116]. We also excluded patients at hospitals with fewer than 300 hospitalizations.

Matching Variables

The variables that we matched are the key variables commonly used for illness severity adjustment: age, sex, laboratory measures on hospital admission, admission diagnoses, comorbid conditions, and admission source. Laboratory values are provided a score in APACHE III-J[109, 113] to account for the non-linear relationships and severity, and thus the scores were used in the analyses. Patients with similar diagnoses were grouped into clinically meaningful categories, as provided in APACHE III-J. We used the comorbidity score from the APACHE.

Profile Matching Procedure

Each hospital was separately used as the focal hospital, and the remaining 112 hospitals served as the control group. The target profile was established as the mean of each variable for the focal hospital. The tolerances were set at 0.05 standard deviations for each covariate, thus ensuring that the standardized mean difference between the groups would be less than 0.1 and therefore balanced[24, 96]. A one-to-one matched sample from the remaining 112 hospitals was found such that the selected matched control group was within the allowable tolerances of the targets. We used a one-to-one matched cohort because we found in the previous simulations in

Chapter 2 that including additional matched controls did not improve estimates. However, we additionally considered a ten-to-one matched cohort as was conducted in the original indirect standardization matching study[68]. We examined how the results changed relative to a one-to-one matched sample as a sensitivity analysis.

Assessing Hospital Quality

After matching, we compared the focal hospital's rate of in-hospital mortality to the rate of in-hospital mortality in the matched control group. We used a doubly robust approach to account for any remaining imbalance between the treated and control groups on key variables. Specifically, we used a multilevel logistic regression model that included a fixed effect with a binary indicator for being in the focal hospital, and adjusted for gender, post-operative status, emergency department admission, age, diagnosis category, comorbidity score, and APACHE score. The model also included a random intercept for the hospital to account for the multilevel nature of the data. We examined whether the indicator for the focal hospital was statistically significant to indicate whether the focal hospital had significantly higher or lower mortality than its matched control group.

Comparison to the Standard Approach of Assessing Hospital Quality

We compared the profile matching assessment to the standard regression-based ranking. We assessed the rate of in-hospital mortality using a multilevel logistic regression model that adjusted for patient-level and hospitalization-level covariates and included a random intercept for the hospital. The model adjusted for the same covariates that were included in the matching algorithm. Hospitals with a significantly positive random intercept had significantly higher mortality than the mean while those with significantly negative random intercepts had significantly lower mortality.

The general form[29] for a multilevel logistic regression model for individual i at cluster j is:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_1 x_{1ij} + \cdots \beta_p x_{pij} + \epsilon_{ij}. \quad (33)$$

For this analysis, we considered the model where p_{ij} was the probability of in-hospital mortality for individual i at hospital j, and the covariates included gender, post-operative status, emergency department admission, age, diagnosis category, comorbidity score, and APACHE score.

Results

There were 125,026 patients at 113 hospitals included. The descriptive characteristics of the patients are provided in Table 2. We computed each hospital's mean for continuous variables or total frequency for binary variables for each variable. The hospital-level median and range of these means are presented in Table 2 to illustrate the spread of the case-mix across hospitals. Notably, the primary diagnosis for patients varied substantially by the hospital. For example, the median hospital rate of cardiovascular diagnosis was 31%, but the hospital with the lowest rate of cardiovascular diagnosis only had 5%, while the highest hospital rate was 89%. There were some hospitals that had no patients with specific diagnoses: gastrointestinal, hematological/ endocrine, metabolic, musculoskeletal/skin, and trauma. The range of hospitals' rates of post-operative patients ranged from 1% to 62%, with the median hospital with 16%. Similarly, there was wide variation in the use of the emergency department as the admission source to the intensive care unit, with hospitals ranging from 3%-83% of all admissions through the emergency department.

Table 2. Descriptive Characteristics of Patients in eICU

Variable	**Overall**	**Hospital Median (IQR)**	**Hospital Range**
Age (every 1 year)	63.0 (17.1)	63.1 (61.5-65.0)	54.5-71.5
Gender (reference: Female)	67883 (54.3%)	54% (52%-57%)	43%-63%
Post-Operative	25410 (20.3%)	16% (10%-25%)	1%-62%
Emergency Department Admission	48910 (39.1%)	43% (30%-57%)	3%-83%
Laboratory & Vital Sign Scores			
Acid Base (pH, PCO2)	0.49 (1.74)	0.49 (0.28-0.66)	0-1.39
Albumin	0.82 (2.50)	0.81 (0.53-1.21)	0-2.99
Bilirubin	0.29 (1.61)	0.27 (0.19-0.36)	0-0.92
Blood urea nitrogen	3.56 (4.29)	3.59 (3.18-4.05)	1.94-5.56
Creatinine	1.54 (2.79)	1.52 (1.36-1.76)	0.64-2.48
Glucose	0.92 (1.82)	0.93 (0.8-1.06)	0.19-1.42
Hematocrit	2.13 (1.36)	2.17 (1.85-2.33)	1.32-2.79
PaO2	1.34 (3.43)	1.03 (0.67-1.62)	0-3.26
Sodium	0.36 (0.82)	0.34 (0.27-0.44)	0.14-0.68
Urine Output	2.50 (4.08)	2.3 (1.26-3.98)	0-6.35
WBC count	0.38 (1.62)	0.37 (0.29-0.45)	0.08-0.77
Heart Rate	3.68 (4.25)	3.74 (3.38-4)	2.09-5.30
Mean Arterial Blood Pressure	10.3 (4.10)	10.27 (9.43-10.84)	7.4-11.92
Respiratory Rate	8.10 (4.76)	7.46 (3.46-13.19)	3.46-13.19
Temperature	0.71 (2.67)	0.55 (0.08-1.87)	0.08-1.87
Glasgow Coma Score (every 1 point)	6.45 (12.6)	4.95 (1.79-17.35)	1.79-17.35
Comorbidity Score (every 1 point)	0.70 (2.87)	0.68 (0.54-0.86)	0.09-2.2
Diagnoses			
Cardiovascular	40162 (32.1%)	31% (25%-38%)	5%-89%
Gastrointestinal	12284 (9.8%)	10% (8%-12%)	0%-19%
Hematological/Endocrine	871 (0.7%)	1% (0%-1%)	0%-2%
Metabolic	10467 (8.4%)	9% (6%-13%)	0%-23%
Musculoskeletal/Skin	1521 (1.2%)	1% (1%-1%)	0%-9%
Neurological	18173 (14.5%)	10% (7%-15%)	1%-66%
Renal/Genitourinary	3098 (2.5%)	3% (2%-3%)	0%-6%
Respiratory	16548 (13.2%)	13% (10%-16%)	3%-32%
Sepsis	16239 (13.0%)	13% (10%-18%)	1%-33%
Trauma	5663 (4.5%)	2% (1%-4%)	0%-23%

Note: Statistics are provided as numbers and percentages for binary variables or mean and standard deviation for continuous variables. IQR=Interquartile range.

For the one-to-one matched cohort, the overall SMD between the focal hospitals and their matched control group ranged from 0.024 to 0.040 across the 113 hospitals for covariates included in the match (median: 0.032). There were 30 (26.5%) hospitals that were significantly different than their matched comparison; 11 (9.7%) hospitals had significantly higher mortality, and 19 (16.8%) hospitals had significantly lower mortality than their matched comparison. With the ten-to-one matched cohort, the overall mean SMD for the covariates included in the matching algorithm ranged from 0.022 to 0.045 across there 113 hospitals (median: 0.040). There were 3 (2.7%) hospitals that had significantly higher mortality, and 11 (9.7%) hospitals had significantly lower mortality. The agreement between the one-to-one and ten-to-one matched cohorts is presented in Figure 15. The Pearson correlation between the estimate using the one-to-one matched sample and the ten-to-one matched sample was 0.877. The main discrepancy between the two is which individual hospitals were significantly different than their comparison.

In Figure 15, the scatter plot depicts the estimate associated with being in the focal hospital using one-to-one matching (x-axis) relative to ten-to-one matching (y-axis). The color indicates whether the effect was significant using one-to-one matching. Green indicates that the hospital under evaluation had a significantly lower mortality rate than its comparator using one-to-one matching. Red indicates that the hospital under evaluation had a significantly higher mortality rate than its comparator using one-to-one matching. Gray and black indicate that there was not a significant difference between the hospital under evaluation and its comparator using one-to-one matching. The shape indicates whether the effect was significant using ten-to-one matching. The triangle indicates that the hospital under evaluation had a significantly lower mortality rate than its comparator using ten-to-one matching. The square denotes significantly higher mortality at the hospital under evaluation using ten-to-one matching. The circle indicates

that there was not a significant difference using ten-to-one matching. The line of best fit (solid) and the line of perfect agreement (dotted) are provided for reference.

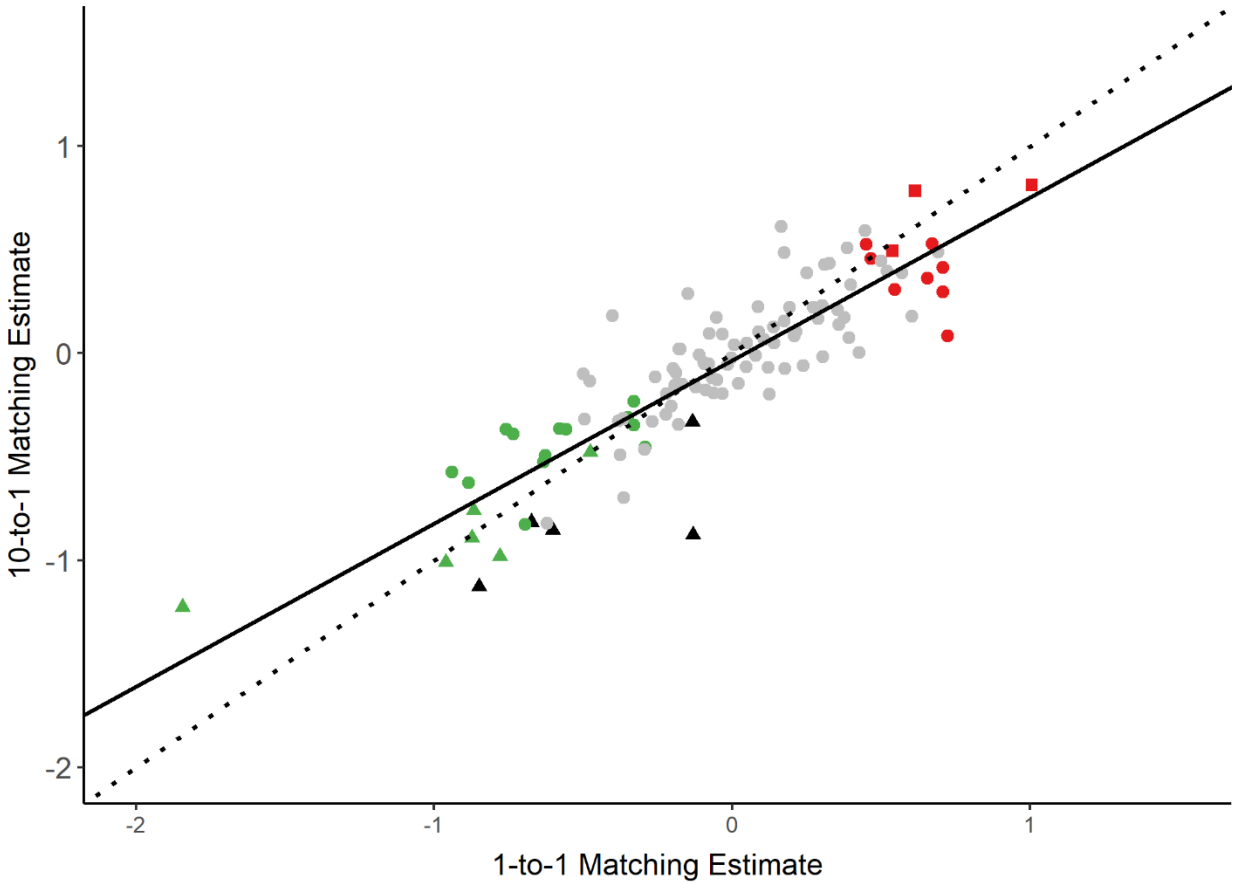


Figure 15. Comparison of One-to-One and Ten-to-One Matched Cohort Hospital Mortality Estimates

Note: The scatter plot depicts the estimate associated with being in the focal hospital using one-to-one matching (x-axis) relative to ten-to-one matching (y-axis).

Table 3. Fixed Effects for the Linear Mixed Effects Model of In-Hospital Mortality

Variable	**Coefficient (95% Cl)**	**OR (95% CI)**
Age (every 1 year)	0.04 (0.03, 0.04)	1.04 (1.03, 1.04)
Gender (reference: Female)	-0.02 (-0.06, 0.03)	0.98 (0.94, 1.03)
Post-Operative	-0.88 (-0.97, -0.8)	0.41 (0.38, 0.45)
Emergency Department Admission	-0.12 (-0.17, -0.07)	0.89 (0.84, 0.93)
Laboratory and Vital Sign Scores		
Acid Base (pH, PCO2)	0.08 (0.07, 0.09)	1.08 (1.07, 1.09)
Albumin	0.06 (0.05, 0.07)	1.06 (1.05, 1.07)
Bilirubin	0.08 (0.07, 0.09)	1.09 (1.07, 1.1)
Blood urea nitrogen	0.02 (0.01, 0.03)	1.02 (1.01, 1.03)
Creatinine	0.07 (0.06, 0.08)	1.07 (1.06, 1.09)
Glucose	0.05 (0.04, 0.06)	1.05 (1.04, 1.06)
Hematocrit	-0.17 (-0.19, -0.15)	0.84 (0.83, 0.86)
PaO2	0.05 (0.04, 0.06)	1.05 (1.04, 1.06)
Sodium	0.08 (0.06, 0.11)	1.09 (1.06, 1.11)
Urine Output	0.01 (0.01, 0.02)	1.01 (1.01, 1.02)
WBC count	0.07 (0.06, 0.08)	1.08 (1.07, 1.09)
Heart Rate	0.08 (0.08, 0.09)	1.08 (1.08, 1.09)
Mean Arterial Blood Pressure	0.05 (0.04, 0.06)	1.05 (1.04, 1.06)
Respiratory Rate	0.01 (0.01, 0.02)	1.01 (1.01, 1.02)
Temperature	0.08 (0.07, 0.09)	1.08 (1.08, 1.09)
Glasgow Coma Score (every 1 point)	0.04 (0.04, 0.04)	1.04 (1.04, 1.04)
Diagnoses (ref: trauma)		
Cardiovascular	-0.19 (-0.32, -0.07)	0.82 (0.73, 0.93)
Gastrointestinal	-0.19 (-0.33, -0.05)	0.83 (0.72, 0.95)
Hematological/Endocrine	-0.03 (-0.31, 0.25)	0.97 (0.73, 1.28)
Metabolic	-1.60 (-1.80, -1.39)	0.20 (0.17, 0.25)
Musculoskeletal/Skin	-0.25 (-0.55, 0.05)	0.78 (0.58, 1.05)
Neurological	0.06 (-0.07, 0.19)	1.07 (0.94, 1.21)
Renal/Genitourinary	-0.63 (-0.84, -0.43)	0.53 (0.43, 0.65)
Respiratory	0.11 (-0.02, 0.24)	1.12 (0.98, 1.27)
Sepsis	-0.04 (-0.17, 0.08)	0.96 (0.84, 1.09)
Comorbidity Score (every 1 point)	0.04 (0.04, 0.05)	1.04 (1.04, 1.05)

Note: OR=Odds Ratio. CI=Confidence Interval. Ref=reference category.

The patient-level fixed effects are provided in Table 3. Overall, of the 113 hospitals, 26 (23.0%) hospitals had a mortality that was significantly higher than the mean, and 20 (17.7%) hospitals had significantly lower mortality. The ICC was 0.035, indicating that 3.5% of the variance in the outcome was attributed to the hospital rather than patient characteristics.

The caterpillar plot depicting the random intercept for the hospital from the multilevel logistic regression model using the entire hospitalization data (the standard approach to hospital benchmarking) is provided in Figure 16. The gray dashed line indicates a random intercept of 0. Hospitals above the line have a higher hospital-level effect of mortality, and hospitals below the line indicate lower hospital-level mortality using the standard regression approach. The colors indicate whether the hospital was identified as having significantly lower (green), higher (red) or no different (black) mortality relative to the matched comparison using indirect standardization profile matching. Overall, there are consistent trends in hospitals with significantly higher or lower mortality rates using indirect standardization profile matching versus the standard regression approach. However, there are fewer hospitals with significant mortality differences when using the larger matched cohort relative to the one-to-one matched cohort.

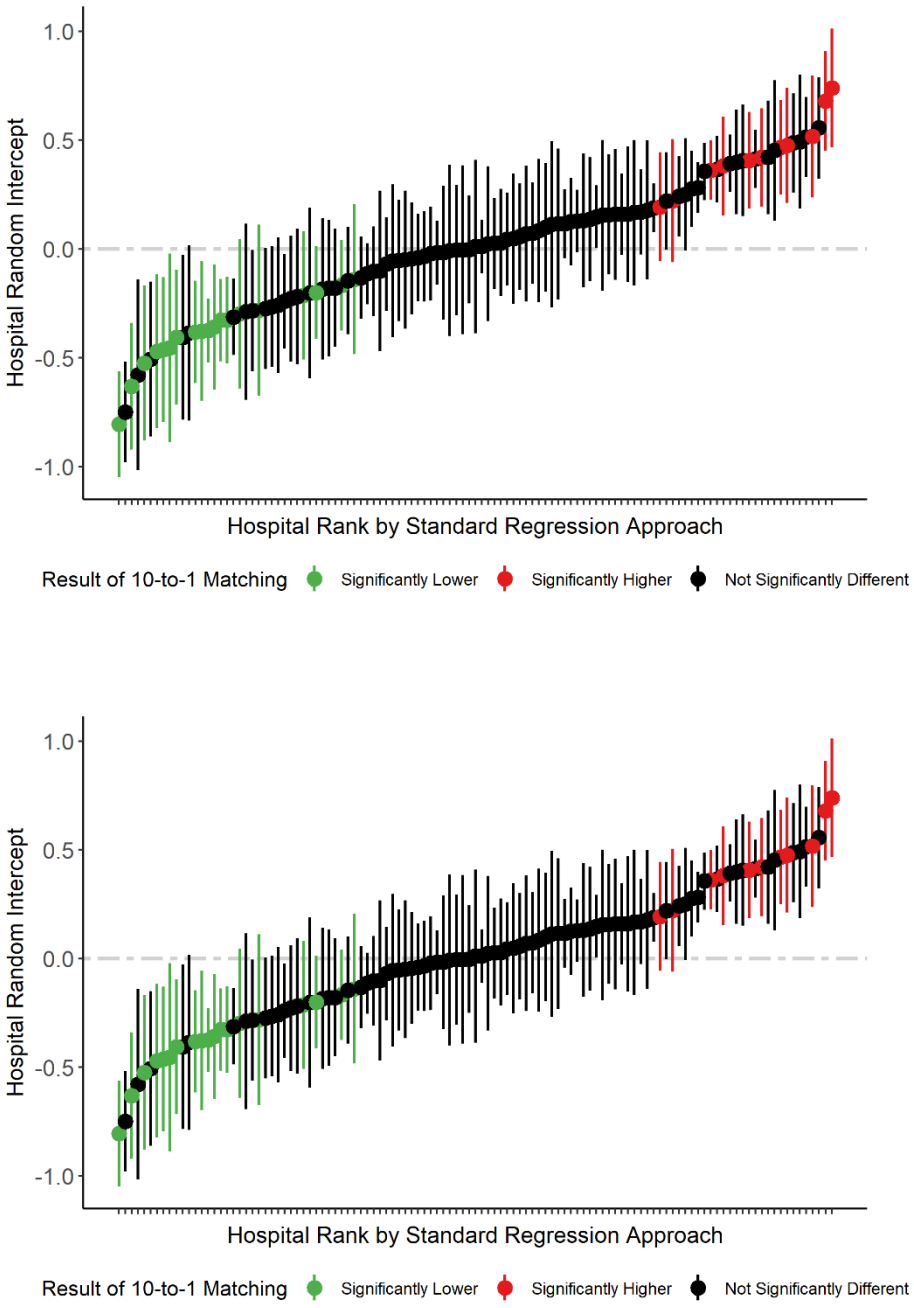


Figure 16. Caterpillar Plot of Hospital Random Intercept Compared to Indirect Standardization Profile Matching

Note: Random intercept (dot) and 95% confidence intervals (vertical bars) for 113 hospitals from a linear mixed effects model of in-hospital mortality with all hospitalizations, adjusted for patient covariates. On the top is the one-to-one and on the bottom is the ten-to-one matched cohorts.

CONCLUSION

In this dissertation, we developed an R package called ProfileMatchit that can be used to employ profile matching. The user can specify a target profile directly and find the largest cardinality match that is balanced with respect to the target profile. In the next chapter, we conducted a simulation study using our package. We tested the impact of increasing the number of matched controls when conducting cardinality matching with multilevel data. In the simulation study, we found that the use of a linear mixed effects model with multilevel matched data was important, even when the ICC was low or if conducting one-to-one matching, to have sufficient coverage. The most important modeling consideration is using a linear mixed effects model to account for the clustering of the control group. There was not a significant benefit to conducting many-to-one matching. However, it is important to consider if the data structure will allow for the appropriate fitting of a multilevel model in this context. If there are problems with model singularity, adding additional matched controls may allow for the estimation of the random effects.

Additionally, we showed that the *a priori* established tolerance level is important because the algorithm finds a solution that meets the tolerance and does not seek to find the *best* solution. Specifically, suppose the tolerance is set at 0.10 standard deviations. In that case, the solution attained may result in a final matched sample that differs by approximately 0.10 standard deviations, even though it may be possible to find a solution that differs by 0.05 standard deviations. We showed in our simulations that this is the case and that the matched sample that differed by 0.10 standard deviations was biased. However, the bias was reduced to a very low level if a doubly robust approach was used. A doubly robust approach adjusted for the remaining

covariate imbalance between the treated and control groups, while the linear mixed effects model accounted for the nesting of the control group.

In the next chapter, we employed profile matching to the setting of hospital quality assessment using a real-world dataset. This application was the culmination of our work to develop an improved version of cardinality matching and provide a new application of profile matching and a better approach to hospital quality assessment.

We proposed a hybrid method that combined profile matching[10] and indirect standardization matching[68] for hospital quality assessment, which we refer to as indirect standardization profile matching (IS-PM). With IS-PM, each patient at the focal hospital is compared to similar patients treated elsewhere. Each hospital in turn serves as the focal hospital, and thus the comparisons are customized to each hospital. The user pre-specified the tolerable differences of each covariate between the treated and control group, and a matched sample is found that is balanced by design. Each hospital is compared to their matched comparison, and the in-hospital mortality rates are assessed using a doubly robust approach, i.e., a linear mixed effects model that is adjusted for covariates. With a doubly robust approach, if either the matching process or the outcome model were correctly specified, the estimated treatment effects will be unbiased.

We found differences in the conclusions about whether a hospital significantly differed in their in-hospital mortality rates when using a one-to-one matched sample relative to a ten-to-one matched sample when conducting IS-PM. However, there was a relationship between the treatment estimates, with a high correlation coefficient ($\rho = 0.88$) between the estimates generated from the one-to-one and ten-to-one matched samples. It may be important to consider

the overall treatment effect estimate when assessing hospital quality rather than whether there was a significant difference.

A major advantage of the proposed approach of IS-PM is that each hospital's comparison is customized to its individual patient case-mix, and thus is a fairer way of assessing patient outcomes. As we showed, there was substantial variation in the distribution of diagnoses between the hospitals. In fact, there were some hospitals that did not treat any patients with some diagnoses. And yet, the traditional regression-based assessment would extrapolate the estimates for those hospitals to the types of patients diagnoses they do not treat. Additionally, the entire patient panel at a given hospital can be included in the assessment, as opposed to hospital-specific template matching which uses a sample of a hospital's patient population to serve as the template by which hospitals are compared. The proposed IS-TM leverages the benefits of matching while also including *all* patients from a given hospital in the assessment.

Limitations and Future Research

A limitation of cardinality matching, and thus profile matching, is that the cardinality matching algorithm does not necessarily find the matched sample that results in the minimum imbalance between the treatment groups but rather finds a solution in which the tolerance is met. Careful consideration should be given to which variables to include in the matching algorithm and the importance of the balance of each variable between the treatment groups and specifying the tolerances for each variable accordingly. Additionally, a doubly robust approach with a linear mixed effects model should be used to address the residual bias. A limitation of the real data study was that it used a publicly available dataset to illustrate the proposed approach and that the sample that we used was not representative of a true hospital system. Further refinement of the

approach may be necessary to extend it to other specific hospital systems based on the specific patient case-mix and variation across hospitals.

An opportunity for future research is to examine the impact of missing data when using cardinality matching and profile matching. Multiple imputation (MI) is one of the most commonly used approaches for the handling of missing data, but there has been little guidance on how best to use MI with matching[117, 118]. There has been some work examining MI with propensity score matching. Granger et al. (2019) and Leyrat et al. (2019) independently showed that MI needs to first impute the propensity scores and then conduct a full MI analysis rather than simply taking the mean of the multiply imputed propensity scores and using them in a single analysis[119, 120]. The use of MI with cardinality matching has not been examined.

There is extensive research on matching for binary treatments[1, 2, 13, 121–123], but generalizations to multiple treatment groups are limited. The most common approach is the use of a generalized propensity score for a multileveled treatment[124–128]. The generalized propensity score is the conditional probability of each individual receiving the treatment condition, given observed covariates[128, 129]. Accounting for all values of treatment in a single model ensures that the treatment effect is estimated for observations that had a non-zero probability of receiving each treatment, thus that the assumption of common support is valid[124, 130]. However, propensity scores generated using multinomial logistic regression are susceptible to extreme propensity scores in the presence of model misspecification[131]. To overcome this limitation, Lopez and Gutman proposed including only those individuals in the region of common support for all treatments[132]. Cardinality matching addresses the issue of common support directly, and thus future research should examine the extension of cardinality matching to the setting of multiple treatment groups.

REFERENCES

[1] STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci* **25** 1–21.

[2] ROSENBAUM, P. R. (2002). *Observational studies*. Springer, New York.

[3] BURN, E., SANCHEZ-SANTOS, M. T., PANDIT, H. G., HAMILTON, T. W., LIDDLE, A. D., MURRAY, D. W. and PINEDO-VILLANUEVA, R. (2018). Ten-year patient-reported outcomes following total and minimally invasive unicompartmental knee arthroplasty: A propensity score-matched cohort analysis. *Knee Surg Sports Traumatol Arthrosc* **26** 1455–64.

[4] DISAIA, P. J., BREWSTER, W. R., ZIOGAS, A. and ANTON-CULVER, H. (2000). Breast cancer survival and hormone replacement therapy: A cohort analysis. *American Journal of Clinical Oncology* **23** 541–5.

[5] CONNORS, J., BASSERI, S., GRANT, A., GIFFIN, N., MAHDI, G., NOBLE, A., RASHID, M., OTLEY, A. and VAN LIMBERGEN, J. (2017). Exclusive enteral nutrition therapy in paediatric Crohn's disease results in long-term avoidance of corticosteroids: Results of a propensity-score matched cohort analysis. *Journal of Crohn's and Colitis* **11** 1063–70.

[6] CHUANG, Y.-C., CHENG, C.-Y., SHENG, W.-H., SUN, H.-Y., WANG, J.-T., CHEN, Y.-C. and CHANG, S.-C. (2014). Effectiveness of tigecycline-based versus colistin- based therapy for treatment of pneumonia caused by multidrug-resistant Acinetobacter baumannii in a critical setting: A matched cohort analysis. *BMC Infect Dis* **14** 102.

[7] COIRO, S., GIRERD, N., ROSSIGNOL, P., FERREIRA, J. P., MAGGIONI, A., PITT, B., TRITTO, I., AMBROSIO, G., DICKSTEIN, K. and ZANNAD, F. (2017). Association of beta-blocker treatment with mortality following myocardial infarction in patients with chronic obstructive pulmonary disease and heart failure or left ventricular dysfunction: A propensity matched-cohort analysis from the high-risk myocardial infarction database initiative. *European Journal of Heart Failure* **19** 271–9.

[8] THOEMMES, F. J. and WEST, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behav Res* **46** 514–43.

[9] 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** 204–31.

[10] COHN, E. R. and ZUBIZARRETA, J. R. (2021). Profile matching for the generalization and personalization of causal inferences. *arXiv:2105.10060 [stat]*.

[11] ZUBIZARRETA, J. R., KILCIOGLU, C. and VIELMA, J. P. (2018). designmatch: Matched samples that are balanced and representative by design. Available at https://CRAN.R-project.org/package=designmatch.

[12] HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* **42** 1–28.

[13] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41.

[14] BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. and STURMER, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology* **163** 1149–56.

[15] RUBIN, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety* **13** 855–7.

[16] SETOGUCHI, S., SCHNEEWEISS, S., BROOKHART, M. A., GLYNN, R. J. and COOK, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety* **17** 546–55.

[17] DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94** 1053–62.

[18] DEHEJIA, R. H. and WAHBA, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* **84** 151–61.

[19] DRAKE, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* **49** 1231–6.

[20] ZHAO, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics* **86** 91–107.

[21] RASSEN, J. A., SHELAT, A. A., MYERS, J., GLYNN, R. J., ROTHMAN, K. J. and SCHNEEWEISS, S. (2012). One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and Drug Safety* **21** 69–80.

[22] AUSTIN, P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology* **172** 1092–7.

[23] MING, K. and ROSENBAUM, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56** 118–24.

[24] LINDEN, A. and SAMUELS, S. J. (2013). Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice* **19** 968–75.

[25] SMITH, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27** 325–53.

[26] VISCONTI, G. and ZUBIZARRETA, J. R. (2018). Handling limited overlap in observational studies with cardinality matching. *Observational Studies* **4** 217–49.

[27] HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15** 199–236.

[28] CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–99.

[29] GELMAN, A. and HILL, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York.

[30] GLYNN, R. J., SCHNEEWEISS, S. and STÜRMER, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & clinical pharmacology & toxicology* **98** 253–9.

[31] DESAI, R. J. and FRANKLIN, J. M. (2019). Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners. *BMJ* **367**.

[32] ZUBIZARRETA, J. R., REINKE, C. E., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2011). Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Am Stat* **65** 229–38.

[33] TRASKIN, M. and SMALL, D. S. (2011). Defining the study population for an observational study to ensure sufficient overlap: A tree approach. *Stat Biosci* **3** 94–118.

[34] ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* **107** 1360–71.

[35] DIAMOND, A. and SEKHON, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics* **95** 932–45.

[36] 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. *J Am Stat Assoc* **110** 515–27.

[37] YU, R., SILBER, J. H. and ROSENBAUM, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statistical Science* **35** 338–55.

[38] IEZZONI, L. I. (1997). The risks of risk adjustment. *JAMA* **278** 1600–7.

[39] PAUL, E., BAILEY, M. and PILCHER, D. (2013). Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: Development

and validation of the Australian and New Zealand risk of death model. *Journal of Critical Care* **28** 935–41.

[40] THE CENTERS FOR MEDICARE & MEDICAID SERVICES. (2021). Overall hospital quality star rating. Available at https://data.cms.gov/provider-data/topics/hospitals/overall-hospital-quality-star-rating/.

[41] MORAN, J. L. and SOLOMON, P. J. (2003). Mortality and other event rates: What do they tell us about performance? *Crit Care Resusc* **5** 292–304.

[42] LILFORD, R., MOHAMMED, M. A., SPIEGELHALTER, D. and THOMSON, R. (2004). Use and misuse of process and outcome data in managing performance of acute medical care: Avoiding institutional stigma. *Lancet* **363** 1147–54.

[43] MOHAMMED, M. A., DEEKS, J. J., GIRLING, A., RUDGE, G., CARMALT, M., STEVENS, A. J. and LILFORD, R. J. (2009). Evidence of methodological bias in hospital standardised mortality ratios: Retrospective database study of English hospitals. *BMJ* **338** b780.

[44] BOTTLE, A., JARMAN, B. and AYLIN, P. (2010). Strengths and weaknesses of hospital standardised mortality ratios. *BMJ* **342** c7116.

[45] LILFORD, R. and PRONOVOST, P. (2010). Using hospital mortality rates to judge hospital performance: A bad idea that just won't go away. *BMJ* **340** c2016.

[46] CASE MIX PROGRAMME. Annual quality report 2018/19 for adult critical care. Available at https://onlinereports.icnarc.org/Reports/2019/12/annual-quality-report-201819-for-adult-critical-care.

[47] CENTERS FOR MEDICARE AND MEDICAID HOSPITAL QUALITY INITIATIVE. Outcome measures. Available at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.

[48] LI, Y.-F. (2018). Strategic analytics for improvement and learning (SAIL): Quality of care. Available at https://www.va.gov/qualityofcare/measure-up/strategic_analytics_for_improvement_and_learning_sail.asp.

[49] ESCOBAR, G. J., GARDNER, M. N., GREENE, J. D., DRAPER, D. and KIPNIS, P. (2013). Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Medical Care* **51** 446–53.

[50] GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series A* **159** 385–443.

[51] DELONG, E. R., PETERSON, E. D., DELONG, D. M., MUHLBAIER, L. H., HACKETT, S. and MARK, D. B. (1997). Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* **16** 2645–64.

[52] NORMAND, S.-L. T. and SHAHIAN, D. M. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science* **22** 206–26.

[53] JONES, H. E. and SPIEGELHALTER, D. J. (2011). The identification of "unusual" health-care providers from a hierarchical model. *The American Statistician* **65** 154–63.

[54] MOHAMMED, M. A., MANKTELOW, B. N. and HOFER, T. P. (2016). Comparison of four methods for deriving hospital standardised mortality ratios from a single hierarchical logistic regression model. *Statistical Methods in Medical Research* **25** 706–15.

[55] ASH, A. S., FIENBERG, S. E., LOUIS, T. A., NORMAND, S.-L. T., STUKEL, T. A. and UTTS, J. (2011). *Statistical issues in assessing hospital performance*. Committee of Presidents of Statistical Societies.

[56] BILIMORIA, K. Y. and BARNARD, C. (2021). An evolving hospital quality star rating system from CMS: Aligning the stars. *JAMA*.

[57] SILBER, J. H., ROSENBAUM, P. R., BRACHET, T. J., ROSS, R. N., BRESSLER, L. J., EVEN-SHOSHAN, O., LORCH, S. A. and VOLPP, K. G. (2010). The hospital compare mortality model and the volume–outcome relationship. *Health Services Research* **45** 1148–67.

[58] HU, W., CHAN, C. W., ZUBIZARRETA, J. R. and ESCOBAR, G. J. (2018). Incorporating longitudinal comorbidity and acute physiology data in template matching for assessing hospital quality: An exploratory study in an integrated health care delivery system. *Medical Care* **56** 448–54.

[59] SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., MUKHERJEE, N., SAYNISCH, P. A., EVEN-SHOSHAN, O., KELZ, R. R. and FLEISHER, L. A. (2014). Template matching for auditing hospital cost and quality. *Health Services Research* **49** 1446–74.

[60] SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., SAYNISCH, P. A., EVEN-SHOSHAN, O., KELZ, R. R. and FLEISHER, L. A. (2014). A hospital-specific template for benchmarking its cost and quality. *Health Services Research* **49** 1475–97.

[61] MOLLING, D., VINCENT, B. M., WIITALA, W. L., ESCOBAR, G. J., HOFER, T. P., LIU, V. X., ROSEN, A. K., RYAN, A. M., SEELYE, S. and PRESCOTT, H. C. (2020). Developing a template matching algorithm for benchmarking hospital performance in a diverse, integrated healthcare system. *Medicine* **99** e20385.

[62] SILBER, J. H., ROSENBAUM, P. R., PIMENTEL, S. D., CALHOUN, S., WANG, W., SHARPE, J. E., REITER, J. G., SHAH, S. A., HOCHMAN, L. L. and EVEN-SHOSHAN, O. (2019). Comparing resource use in medical admissions of children with complex chronic conditions. *Medical Care* **57** 615–24.

[63] VINCENT, B. M., WIITALA, W. L., LUGINBILL, K. A., MOLLING, D. J., HOFER, T. P., RYAN, A. M. and PRESCOTT, H. C. (2019). Template matching for benchmarking hospital performance in the Veterans Affairs healthcare system. *Medicine (Baltimore)* **98** e15644.

[64] VINCENT, B. M., MOLLING, D., ESCOBAR, G. J., HOFER, T. P., IWASHYNA, T. J., LIU, V. X., ROSEN, A. K., RYAN, A. M., SEELYE, S., WIITALA, W. L. and PRESCOTT, H. C. (2021). Hospital-specific template matching for benchmarking performance in a diverse multihospital system. *Medical Care* **59** 1090–8.

[65] SHAHIAN, D. M. (2020). Making the case for teaching hospitals: Evolving metrics and methodologies. *Annals of Surgery* **271** 422–4.

[66] SILBER, J. H., ROSENBAUM, P. R., NIKNAM, B. A., ROSS, R. N., REITER, J. G., HILL, A. S., HOCHMAN, L. L., BROWN, S. E., ARRIAGA, A. F., KELZ, R. R. and FLEISHER, L. A. (2020). Comparing outcomes and costs of surgical patients treated at major teaching and nonteaching hospitals: A national matched analysis. *Annals of Surgery* **271** 412–21.

[67] SILBER, J. H., ROSENBAUM, P. R., MCHUGH, M. D., LUDWIG, J. M., SMITH, H. L., NIKNAM, B. A., EVEN-SHOSHAN, O., FLEISHER, L. A., KELZ, R. R. and AIKEN, L. H. (2016). Comparing the value of better nursing work environments across different levels of patient risk. *JAMA Surg* **151** 527–36.

[68] SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., HILL, A. S., EVEN-SHOSHAN, O., KELZ, R. R. and FLEISHER, L. A. (2016). Indirect standardization matching: Assessing specific advantage and risk synergy. *Health Services Research* **51** 2330–57.

[69] SILBER, J. H., ROSENBAUM, P. R., WANG, W., LUDWIG, J. M., CALHOUN, S., GUEVARA, J. P., ZORC, J. J., ZEIGLER, A. and EVEN-SHOSHAN, O. (2016). Auditing practice style variation in pediatric inpatient asthma care. *JAMA Pediatrics* **170** 878–86.

[70] IMBENS, G. W. and RUBIN, D. B. (2010). *Rubin causal model*. S. N. Durlauf and L. E. Blume, ed Palgrave Macmillan UK, London.

[71] HOLLAND, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81** 945–60.

[72] RUBIN, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press, Cambridge.

[73] DING, P. and LI, F. (2018). Causal inference: A missing data perspective. *Statist. Sci.* **33**.

[74] ROSENBAUM, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association* **79** 41–8.

[75] RUBIN, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics* **36** 293–8.

[76] DENG, A. (2021). Causal inference and its applications in online industry. Available at https://alexdeng.github.io/causal/.

[77] FUNK, M. J., WESTREICH, D., WIESEN, C., STÜRMER, T., BROOKHART, M. A. and DAVIDIAN, M. (2011). Doubly robust estimation of causal effects. *Am J Epidemiol* **173** 761–7.

[78] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–66.

[79] BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–73.

[80] YANG, S. (2017). Propensity score weighting for causal inference with clustered data. *arXiv:1703.06086 [stat]*.

[81] LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23** 2937–60.

[82] HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal inference: What if*. Chapman & Hall/ CRC, Bota Raton.

[83] ALVES, M. F. (2021). Causal inference for the brave and true. Available at https://matheusfacure.github.io/python-causality-handbook/12-Doubly-Robust-Estimation.html.

[84] BERTSIMAS, D. and TSITSIKLIS, J. (1997). *Introduction to linear optimization*. Athena Scientific.

[85] ANON. (2021). *gurobi: Gurobi optimizer 9.1 interface*. Gurobi Optimization LLC.

[86] ANAND, R., AGGARWAL, D. and KUMAR, V. (2017). A comparative analysis of optimization solvers. *Journal of Statistics and Management Systems* **20** 623–35.

[87] 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** 628–36.

[88] BARRATT, S., ANGERIS, G. and BOYD, S. (2020). Optimal representative sample weighting. *arXiv: 2005.09065 [stat.ML]*.

[89] WINSHIP, C. and RADBILL, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research* **23** 230–57.

[90] WESTREICH, D., LESSLER, J. and FUNK, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* **63** 826–33.

[91] ABRAHAM, B. and LEDOLTER, J. (2006). *Introduction to regression modeling*. Brooks/Cole Cengage Learning.

[92] FINCH, W. H., BOLIN, J. E. and KELLEY, K. (2019). *Multilevel modeling using R*. CRC Press/Taylor & Francis, New York, NY.

[93] CHUANG, J.-H., HRIPCSAK, G. and HEITJAN, D. F. (2002). Design and analysis of controlled trials in naturally clustered environments: Implications for medical informatics. *Journal of the American Medical Informatics Association* **9** 230–8.

[94] UKOUMUNNE, O., GULLIFORD, M., CHINN, S., STERNE, J. and BURNEY, P. (1999). Methods for evaluating area-wide and organisation-based interventions in health and health care: A systematic review. *Health Technology Assessment* **3** iii–92.

[95] ZEGER, S. L. and LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 121–30.

[96] NORMAND, S. T., LANDRUM, M. B., GUADAGNOLI, E., AYANIAN, J. Z., RYAN, T. J., CLEARY, P. D. and MCNEIL, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *J Clin Epidemiol* **54** 387–98.

[97] RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology* **2** 169–88.

[98] AUSTIN, P. C. (2008). Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiol Drug Saf* **17** 1218–25.

[99] 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** 744–57.

[100] COOKE, C. R., KENNEDY, E. H., WIITALA, W. L., ALMENOFF, P. L., SALES, A. E. and IWASHYNA, T. J. (2012). Despite variation in volume, Veterans Affairs hospitals show consistent outcomes among patients with non-postoperative mechanical ventilation. *Critical Care Medicine* **40** 2569–75.

[101] PRESCOTT, H. C. (2017). Variation in postsepsis readmission patterns: A cohort study of Veterans Affairs beneficiaries. *Annals of the American Thoracic Society* **14** 230–7.

[102] VIGLIANTI, E. M., BAGSHAW, S. M., BELLOMO, R., MCPEAKE, J., WANG, X. Q., SEELYE, S. and IWASHYNA, T. J. (2020). Hospital-level variation in the development of persistent critical illness. *Intensive Care Med* **46** 1567–75.

[103] HUA, M. J. and FEINGLASS, J. (2022). Variations in COVID-19 hospital mortality by patient race/ethnicity and hospital type in Illinois. *J. Racial and Ethnic Health Disparities*.

[104] GUROBI OPTIMIZATION, LLC. (2021). Gurobi optimizer reference manual.

[105] CASELLA, G., Berger, Roger L. (2021). *Statistical inference*. Brooks/Cole Cengage Learning, Belmont (California).

[106] NIKNAM, B. A. and ZUBIZARRETA, J. R. (2022). Using cardinality matching to design balanced and representative samples for observational studies. *JAMA* **327** 173–4.

[107] GOLDBERGER, A. L., AMARAL, L. A., GLASS, L., HAUSDORFF, J. M., IVANOV, P. C., MARK, R. G., MIETUS, J. E., MOODY, G. B., PENG, C. K. and STANLEY, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101** E215-220.

[108] POLLARD, T. J., JOHNSON, A. E. W., RAFFA, J. D., CELI, L. A., MARK, R. G. and BADAWI, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data* **5** 180178.

[109] ZIMMERMAN, J. E., KRAMER, A. A., MCNAIR, D. S. and MALILA, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine* **34**.

[110] HOSMER, D. W. and LEMESHOW, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **9** 1043–69.

[111] TEASDALE, G. and JENNETT, B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet* **2** 81–4.

[112] SIMKINS, M., IQBAL, A., GRONEMEYER, A., KONZEN, L., WHITE, J., KOENIG, M., PALMER, C., KERBY, P., BUCKMAN, S., DESPOTOVIC, V., HOEHNER, C. and BOYLE, W. (2019). Inter-rater reliability and impact of disagreements on acute physiology and chronic health evaluation IV mortality predictions. *Critical Care Explorations* **1** e0059.

[113] KNAUS, W. A., WAGNER, D. P., DRAPER, E. A., ZIMMERMAN, J. E., BERGNER, M., BASTOS, P. G., SIRIO, C. A., MURPHY, D. J., LOTRING, T., DAMIANO, A. and HARRELL, F. E. (1991). The APACHE III prognostic system. *Chest* **100** 1619–36.

[114] ANZICS CORE. (2021). APD data dictionary for software programmers.ANZICS Centre for Outcome and Resource Evaluation. Available at https://www.anzics.com.au/wp-content/uploads/2018/08/ANZICS-APD-Dictionary-Programmers.pdf.

[115] FENG, S. and DUBIN, J. A. (2021). Identifying early-measured variables associated with APACHE IVa providing incorrect in-hospital mortality predictions for critical care patients. *Sci Rep* **11** 22203.

[116] COSGRIFF, C. V., CELI, L. A., KO, S., SUNDARESAN, T., ARMENGOL DE LA HOZ, M. Á., KAUFMAN, A. R., STONE, D. J., BADAWI, O. and DELIBERATO, R. O. (2019). Developing well-calibrated illness severity scores for decision support in the critically ill. *npj Digit. Med.* **2** 1–8.

[117] RUBIN, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

[118] SCHAFER, J. L. (1999). Multiple imputation: A primer. *Statistical methods in medical research* **8** 3–15.

[119] GRANGER, E., SERGEANT, J. C. and LUNT, M. (2019). Avoiding pitfalls when combining multiple imputation and propensity scores. *Statistics in Medicine* **38** 5120–32.

[120] LEYRAT, C., CAILLE, A., FOUCHER, Y. and GIRAUDEAU, B. (2016). Propensity score to detect baseline imbalance in cluster randomized trials: The role of the C-statistic. *BMC Medical Research Methodology* **16** 9–9.

[121] COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya-the Indian Journal of Statistics Series A* **35** 417–46.

[122] RUBIN, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.

[123] AUSTIN, P. C. (2008). The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology* **61** 537–45.

[124] GARRIDO, M. M., LUM, J. and PIZER, S. D. (2021). Vector-based kernel weighting: A simple estimator for improving precision and bias of average treatment effects in multiple treatment settings. *Statistics in Medicine* **40** 1204–23.

[125] LI, F., ZASLAVSKY, A. M. and LANDRUM, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine* **32** 3373–87.

[126] YOSHIDA, K., HERNÁNDEZ-DÍAZ, S., SOLOMON, D. H., JACKSON, J. W., GAGNE, J. J., GLYNN, R. J. and FRANKLIN, J. M. (2017). Matching weights to simultaneously compare three treatment groups. *Epidemiology* **28** 387–95.

[127] KOH, W. Y. and TU, C. (2021). A hybrid generalized propensity score approach for observational studies. *Communications in Statistics - Simulation and Computation* **0** 1–10.

[128] IMBENS, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–10.

[129] LEITE, W. (2016). *Practical propensity score methods using R*. Sage Publications.

[130] RASSEN, J. A., SHELAT, A. A., FRANKLIN, J. M., GLYNN, R. J., SOLOMON, D. H. and SCHNEEWEISS, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24** 401–9.

[131] LIN, L., ZHU, Y. and CHEN, L. (2019). Causal inference for multi-level treatments with machine-learned propensity scores. *Health Serv Outcomes Res Method* **19** 106–26.

[132] LOPEZ, M. J. and GUTMAN, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science* **32** 432–54, 23.

[133] GUPTA, B., JAIN, G., CHANDRAKAR, S., GUPTA, N. and AGARWAL, A. (2021). Arterial blood gas as a predictor of mortality in covid pneumonia patients initiated on noninvasive mechanical ventilation: A retrospective analysis. *Indian J Crit Care Med* **25** 866–71.

APPENDIX A. SUPPLEMENTAL TABLES

Table A1. Mean Standardized Mean Difference Across the Measured Covariates

Parameters		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	0.300	0.047	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.3	0.05	0.300	0.047	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.3	0.10	0.300	0.047	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.4	0.03	0.400	0.048	0.049	0.049	0.049	0.050	0.050	0.050	0.049	0.049	0.049	0.050
0.4	0.05	0.400	0.048	0.049	0.049	0.049	0.050	0.050	0.050	0.049	0.049	0.049	0.050
0.4	0.10	0.400	0.048	0.049	0.049	0.049	0.050	0.050	0.050	0.049	0.049	0.049	0.050
0.5	0.03	0.500	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.5	0.05	0.500	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.5	0.10	0.500	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
Linear Mixed Effects Model													
0.3	0.03	0.300	0.047	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.3	0.05	0.300	0.047	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.3	0.10	0.300	0.047	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.4	0.03	0.400	0.048	0.049	0.049	0.049	0.050	0.050	0.050	0.049	0.049	0.049	0.050
0.4	0.05	0.400	0.048	0.049	0.049	0.049	0.050	0.050	0.050	0.049	0.049	0.049	0.050
0.4	0.10	0.400	0.048	0.049	0.049	0.049	0.050	0.050	0.050	0.049	0.049	0.049	0.050
0.5	0.03	0.500	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.5	0.05	0.500	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050
0.5	0.10	0.500	0.048	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset.

Table A2. Bias of the Treatment Estimate for Unadjusted Models

Parameters		Number of Matched Controls											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	8.38	1.26	1.31	1.34	1.33	1.34	1.34	1.33	1.23	1.18	1.21	1.37
0.3	0.05	8.37	1.24	1.29	1.32	1.32	1.33	1.32	1.32	1.22	1.17	1.19	1.36
0.3	0.10	8.48	1.37	1.41	1.45	1.44	1.45	1.45	1.44	1.34	1.29	1.31	1.48
0.4	0.03	11.18	1.28	1.33	1.34	1.28	1.30	1.28	1.27	1.05	1.11	1.16	1.38
0.4	0.05	11.17	1.28	1.32	1.34	1.27	1.29	1.27	1.26	1.04	1.11	1.15	1.37
0.4	0.10	11.15	1.26	1.30	1.32	1.25	1.27	1.25	1.25	1.02	1.09	1.13	1.35
0.5	0.03	13.98	1.30	1.33	1.33	1.21	1.26	1.24	1.24	1.16	1.17	1.22	1.37
0.5	0.05	13.97	1.30	1.32	1.32	1.20	1.25	1.24	1.24	1.15	1.17	1.21	1.37
0.5	0.10	13.96	1.29	1.31	1.31	1.19	1.24	1.22	1.22	1.14	1.15	1.20	1.36
Linear Mixed Effects Model													
0.3	0.03	8.38	1.25	1.31	1.34	1.33	1.34	1.34	1.33	1.24	1.18	1.21	1.37
0.3	0.05	8.37	1.24	1.30	1.32	1.32	1.33	1.32	1.32	1.22	1.17	1.19	1.36
0.3	0.10	8.48	1.37	1.42	1.45	1.44	1.45	1.45	1.44	1.34	1.29	1.31	1.48
0.4	0.03	11.18	1.28	1.33	1.34	1.28	1.30	1.28	1.27	1.05	1.12	1.16	1.38
0.4	0.05	11.17	1.28	1.32	1.34	1.27	1.30	1.27	1.26	1.04	1.11	1.15	1.37
0.4	0.10	11.15	1.26	1.30	1.32	1.26	1.28	1.25	1.25	1.02	1.09	1.13	1.35
0.5	0.03	13.98	1.30	1.33	1.33	1.21	1.26	1.24	1.24	1.16	1.17	1.22	1.37
0.5	0.05	13.97	1.30	1.32	1.32	1.20	1.25	1.24	1.24	1.16	1.17	1.21	1.37
0.5	0.10	13.96	1.28	1.31	1.31	1.19	1.24	1.22	1.22	1.14	1.15	1.20	1.36

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. The bias is the difference between the estimated treatment effect and the true treatment effect (0). Presented is the mean of the bias of the treatment effect across the 1,000 simulations.

Table A3. Coverage of the Treatment Effect Estimate 95% Confidence Interval for Unadjusted Models

Parameters		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	0.000	0.314	0.256	0.236	0.221	0.216	0.207	0.204	0.203	0.210	0.203	0.143
0.3	0.05	0.001	0.236	0.178	0.161	0.153	0.148	0.143	0.140	0.155	0.136	0.150	0.106
0.3	0.10	0.007	0.225	0.194	0.175	0.169	0.166	0.155	0.159	0.150	0.164	0.156	0.112
0.4	0.03	0.000	0.293	0.244	0.220	0.224	0.214	0.212	0.208	0.209	0.196	0.200	0.127
0.4	0.05	0.000	0.269	0.218	0.191	0.196	0.180	0.180	0.174	0.183	0.173	0.171	0.128
0.4	0.10	0.001	0.177	0.142	0.128	0.124	0.120	0.116	0.111	0.126	0.103	0.115	0.070
0.5	0.03	0.000	0.291	0.233	0.210	0.204	0.199	0.180	0.167	0.184	0.170	0.164	0.123
0.5	0.05	0.000	0.258	0.211	0.188	0.186	0.175	0.159	0.154	0.167	0.147	0.141	0.121
0.5	0.10	0.000	0.216	0.179	0.164	0.152	0.152	0.135	0.126	0.124	0.130	0.124	0.094
Linear Mixed Effects Model													
0.3	0.03	0.002	0.927	0.891	0.884	0.883	0.877	0.879	0.881	0.887	0.898	0.896	0.868
0.3	0.05	0.034	0.895	0.850	0.794	0.780	0.762	0.747	0.743	0.758	0.759	0.744	0.724
0.3	0.10	0.231	0.946	0.922	0.926	0.925	0.926	0.923	0.920	0.924	0.934	0.918	0.918
0.4	0.03	0.000	0.907	0.886	0.872	0.872	0.880	0.882	0.873	0.892	0.894	0.893	0.872
0.4	0.05	0.002	0.932	0.907	0.905	0.912	0.906	0.909	0.909	0.915	0.918	0.914	0.905
0.4	0.10	0.046	0.895	0.832	0.801	0.774	0.766	0.750	0.742	0.777	0.753	0.747	0.737
0.5	0.03	0.000	0.863	0.851	0.861	0.878	0.877	0.870	0.877	0.889	0.891	0.891	0.872
0.5	0.05	0.000	0.900	0.900	0.898	0.911	0.906	0.909	0.912	0.915	0.913	0.912	0.904
0.5	0.10	0.011	0.935	0.920	0.921	0.928	0.929	0.927	0.925	0.927	0.928	0.930	0.922

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. The coverage is the proportion of the 1,000 simulations in which the true treatment effect is included in the 95% confidence interval for the estimated treatment effect. Coverage is much higher when using a linear mixed effects model than when using a linear model.

Table A4. Mean Squared Error of the Treatment Estimate for Unadjusted Models

Parameter		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	93.1	5.7	5.9	5.9	5.9	6.0	6.0	5.9	5.7	5.5	5.6	6.1
0.3	0.05	95.4	8.2	8.3	8.4	8.4	8.4	8.4	8.4	8.2	7.9	8.0	8.5
0.3	0.10	105.2	15.9	16.0	16.0	16.1	16.1	16.1	16.0	15.7	15.5	15.6	16.2
0.4	0.03	162.9	5.8	5.9	6.0	5.8	5.9	5.8	5.8	5.2	5.3	5.4	6.1
0.4	0.05	165.2	8.4	8.4	8.6	8.3	8.5	8.3	8.4	7.7	7.9	8.0	8.6
0.4	0.10	171.3	15.1	15.1	15.3	15.0	15.2	15.0	15.1	14.4	14.6	14.7	15.3
0.5	0.03	252.6	5.8	5.9	5.9	5.5	5.7	5.8	5.7	5.4	5.5	5.6	6.1
0.5	0.05	254.9	8.4	8.5	8.5	8.1	8.3	8.4	8.3	7.9	8.0	8.1	8.6
0.5	0.10	261.2	15.3	15.3	15.4	14.9	15.2	15.3	15.2	14.8	14.9	15.0	15.5
Linear Mixed Effects Model													
0.3	0.03	93.1	5.7	5.9	5.9	5.9	6.0	6.0	5.9	5.7	5.5	5.6	6.1
0.3	0.05	95.4	8.2	8.3	8.4	8.4	8.4	8.4	8.3	8.2	7.9	8.0	8.5
0.3	0.10	105.2	15.8	16.0	16.0	16.1	16.1	16.1	16.0	15.7	15.5	15.6	16.2
0.4	0.03	162.9	5.8	5.9	6.0	5.8	5.9	5.8	5.8	5.2	5.3	5.4	6.1
0.4	0.05	165.2	8.4	8.4	8.6	8.3	8.5	8.4	8.4	7.7	7.9	8.0	8.6
0.4	0.10	171.4	15.1	15.1	15.3	14.9	15.2	15.1	15.1	14.4	14.6	14.7	15.3
0.5	0.03	252.6	5.8	5.9	5.9	5.5	5.7	5.8	5.7	5.4	5.5	5.6	6.1
0.5	0.05	254.9	8.4	8.5	8.5	8.1	8.3	8.4	8.3	7.9	8.0	8.1	8.6
0.5	0.10	261.2	15.3	15.4	15.4	15.0	15.2	15.3	15.2	14.8	14.9	15.0	15.5

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. The means squared error is a measure of precision of an estimator, and is defined as the sum of the variance of the estimator and the square of the bias of the estimator across the 1,000 simulations.

Table A5. Variance of the Treatment Effect Estimate for Unadjusted Models

Parameter		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	23.0	4.1	4.2	4.2	4.2	4.2	4.2	4.1	4.2	4.1	4.1	4.2
0.3	0.05	25.4	6.6	6.7	6.6	6.7	6.7	6.7	6.6	6.7	6.6	6.6	6.7
0.3	0.10	33.3	14.0	14.0	13.9	14.0	14.0	14.0	13.9	13.9	13.8	13.9	14.0
0.4	0.03	38.0	4.2	4.1	4.2	4.1	4.2	4.2	4.2	4.1	4.1	4.1	4.2
0.4	0.05	40.4	6.8	6.7	6.8	6.7	6.8	6.7	6.8	6.6	6.7	6.7	6.8
0.4	0.10	47.0	13.5	13.4	13.5	13.4	13.5	13.5	13.6	13.3	13.4	13.4	13.5
0.5	0.03	57.3	4.1	4.1	4.2	4.1	4.1	4.2	4.2	4.0	4.1	4.1	4.2
0.5	0.05	59.7	6.7	6.7	6.7	6.6	6.7	6.8	6.8	6.6	6.6	6.6	6.8
0.5	0.10	66.4	13.7	13.6	13.7	13.5	13.6	13.8	13.7	13.5	13.6	13.6	13.7
Linear Mixed Effects Model													
0.3	0.03	23.0	4.1	4.2	4.2	4.2	4.2	4.2	4.1	4.2	4.1	4.1	4.2
0.3	0.05	25.4	6.6	6.7	6.6	6.7	6.7	6.7	6.6	6.7	6.6	6.6	6.7
0.3	0.10	33.3	14.0	14.0	13.9	14.0	14.0	14.0	13.9	13.9	13.8	13.9	14.0
0.4	0.03	38.0	4.2	4.1	4.2	4.1	4.2	4.2	4.2	4.1	4.1	4.1	4.2
0.4	0.05	40.4	6.7	6.7	6.8	6.7	6.8	6.7	6.8	6.6	6.7	6.7	6.8
0.4	0.10	47.0	13.5	13.4	13.5	13.4	13.6	13.5	13.6	13.4	13.4	13.4	13.5
0.5	0.03	57.3	4.1	4.1	4.2	4.1	4.1	4.2	4.2	4.0	4.1	4.1	4.2
0.5	0.05	59.7	6.7	6.7	6.8	6.6	6.7	6.8	6.8	6.6	6.6	6.6	6.8
0.5	0.10	66.4	13.7	13.7	13.7	13.5	13.7	13.8	13.7	13.5	13.6	13.6	13.7

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. Reported are the variance of the treatment effect estimates across the 1,000 simulations. The sampling variance was only marginally impacted by the number of matched controls.

Table A6. Bias of the Treatment Estimate for Covariate-Adjusted Models

Parameters		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02
0.3	0.05	-0.03	-0.04	-0.04	-0.03	-0.04	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.03
0.3	0.10	0.09	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.08	0.09	0.09	0.09
0.4	0.03	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
0.4	0.05	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
0.4	0.10	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
0.5	0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
0.5	0.05	-0.03	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
0.5	0.10	-0.04	-0.04	-0.05	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
Linear Mixed Effects Model													
0.3	0.03	-0.02	-0.03	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
0.3	0.05	-0.03	-0.04	-0.03	-0.03	-0.03	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
0.3	0.10	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
0.4	0.03	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
0.4	0.05	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
0.4	0.10	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
0.5	0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
0.5	0.05	-0.03	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
0.5	0.10	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. The bias is the difference between the estimated treatment effect and the true treatment effect (0). Presented is the mean of the bias of the treatment effect across the 1,000 simulations.

Table A7. Coverage of the Treatment Effect Estimate 95% Confidence Interval for Covariate-Adjusted Models

Parameters		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	0.108	0.148	0.131	0.124	0.121	0.121	0.122	0.126	0.118	0.116	0.112	0.113
0.3	0.05	0.254	0.279	0.278	0.274	0.259	0.260	0.261	0.263	0.259	0.255	0.260	0.253
0.3	0.10	0.067	0.073	0.073	0.077	0.070	0.068	0.071	0.069	0.070	0.066	0.071	0.068
0.4	0.03	0.111	0.144	0.130	0.122	0.118	0.115	0.120	0.115	0.114	0.116	0.113	0.114
0.4	0.05	0.085	0.114	0.105	0.099	0.097	0.092	0.089	0.093	0.090	0.093	0.091	0.087
0.4	0.10	0.250	0.251	0.252	0.251	0.252	0.249	0.243	0.246	0.247	0.250	0.253	0.251
0.5	0.03	0.113	0.149	0.137	0.128	0.120	0.115	0.118	0.112	0.110	0.115	0.113	0.113
0.5	0.05	0.086	0.117	0.101	0.100	0.098	0.091	0.091	0.087	0.086	0.091	0.090	0.087
0.5	0.10	0.075	0.086	0.076	0.072	0.076	0.073	0.076	0.073	0.074	0.075	0.073	0.073
Linear Mixed Effects Model													
0.3	0.03	0.945	0.942	0.946	0.942	0.942	0.944	0.943	0.942	0.941	0.942	0.942	0.942
0.3	0.05	0.938	0.942	0.945	0.944	0.941	0.943	0.940	0.941	0.938	0.940	0.941	0.941
0.3	0.10	0.943	0.942	0.943	0.945	0.944	0.946	0.943	0.945	0.944	0.944	0.942	0.942
0.4	0.03	0.945	0.938	0.942	0.944	0.941	0.945	0.944	0.943	0.940	0.944	0.941	0.941
0.4	0.05	0.943	0.941	0.942	0.943	0.944	0.944	0.944	0.943	0.939	0.940	0.942	0.941
0.4	0.10	0.938	0.941	0.943	0.939	0.940	0.941	0.937	0.939	0.941	0.940	0.944	0.942
0.5	0.03	0.945	0.946	0.941	0.943	0.944	0.944	0.945	0.942	0.945	0.942	0.941	0.941
0.5	0.05	0.943	0.944	0.942	0.942	0.943	0.945	0.945	0.942	0.943	0.943	0.944	0.941
0.5	0.10	0.943	0.945	0.941	0.944	0.945	0.945	0.944	0.943	0.943	0.942	0.942	0.942

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. The coverage is the proportion of the 1,000 simulations in which the true treatment effect is included in the 95% confidence interval for the estimated treatment effect. Coverage is much higher when using a linear mixed effects model than when using a linear model.

Table A8. Mean Squared Error of the Treatment Estimate for Covariate-Adjusted Models

Parameter		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.3	0.05	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2
0.3	0.10	13.4	13.5	13.5	13.5	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4
0.4	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.4	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.4	0.10	13.1	13.1	13.1	13.1	13.1	13.0	13.0	13.1	13.1	13.1	13.1	13.1
0.5	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.5	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.5	0.10	13.3	13.3	13.2	13.3	13.2	13.2	13.3	13.3	13.3	13.3	13.3	13.3
Linear Mixed Effects Model													
0.3	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.3	0.05	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2
0.3	0.10	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4
0.4	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.4	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.4	0.10	13.1	13.1	13.0	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1
0.5	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.5	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.5	0.10	13.3	13.2	13.2	13.3	13.2	13.3	13.3	13.3	13.3	13.3	13.3	13.3

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. The means square error is a measure of precision of an estimator, and is defined as the sum of the variance of the estimator and the square of the bias of the estimator across the 1,000 simulations.

Table A9. Variance of the Treatment Effect Estimate for Covariate-Adjusted Models

Parameter		**Number of Matched Controls**											
SMD	ICC	0	1	2	3	4	5	6	7	8	9	10	Inf
Linear Model													
0.3	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.3	0.05	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2
0.3	0.10	13.4	13.5	13.4	13.5	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4
0.4	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.4	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.4	0.10	13.1	13.1	13.1	13.1	13.1	13.0	13.0	13.1	13.1	13.1	13.1	13.1
0.5	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.5	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.5	0.10	13.3	13.3	13.2	13.3	13.2	13.2	13.3	13.3	13.3	13.3	13.3	13.3
Linear Mixed Effects Model													
0.3	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.3	0.05	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2
0.3	0.10	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4	13.4
0.4	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.4	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.4	0.10	13.1	13.1	13.0	13.1	13.1	13.1	13.0	13.1	13.1	13.1	13.0	13.1
0.5	0.03	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
0.5	0.05	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
0.5	0.10	13.3	13.2	13.2	13.3	13.2	13.3	13.3	13.3	13.3	13.3	13.3	13.3

Note: SMD=standardized mean difference. ICC=intraclass correlation coefficient. Inf indicates largest subset. Reported are the variance of the treatment effect estimates across the 1,000 simulations. The sampling variance was only marginally impacted by the number of matched controls.

Table A10. APACHE Points for the Acute Physiology Score

APS Variable	**APS Points**
Heart Rate (beats per minute)	≤39: 8 40-49: 5 50-99: 0 100-109: 1 110-119: 5 120-139: 7 140-154: 13 ≥155: 17
Mean Arterial Blood Pressure (mmHg)	≤39: 23 40-59: 15 60-69: 7 70-79: 6 80-99: 0 100-119: 4 120-129: 7 130-139: 9 ≥140: 10
Temperature (Celsius)	≤32.9: 20 33.0-33.4: 16 33.4-33.9: 13 34.0-34.9: 8 35.0-35.9: 2 36.0-39.9: 0 ≥40.0: 4
Respiratory Rate (breaths per minute)	≤5: 17 6-11: 8 12-13: 7 14-24: 0 25-34: 6 35-39: 9 40-49: 11 ≥50: 18 For ventilated patients, if respiratory rate 6-12: 0

Table A10. APACHE Points for the Acute Physiology Score (continued)

APS Variable	**APS Points**
Acid Base (pH, pCO2)	pH < 7.20 & pCO2< 50: 12 pH < 7.20: 4 pH < 7.30 & pCO2< 30: 9 pH < 7.30 & pCO2< 40: 6 pH < 7.30 & pCO2<50: 3 pH < 7.30 & pCO2≥ 50: 2 pH < 7.35 & pCO2< 30: 9 pH < 7.35 & pCO2< 45: 0 pH < 7.35 & pCO2≥45: 1 pH < 7.45 & pCO2<30: 5 pH < 7.45 & pCO2< 45: 0 pH < 7.45 & pCO2≥45: 1 pH < 7.50 & pCO2< 30: 5 pH < 7.50 & pCO2< 35: 0 pH < 7.50 & pCO2< 45: 2 pH < 7.50 & pCO2≥ 45: 12 pH < 7.60 & pCO2< 40: 3 pH < 7.60 & pCO2≥40: 12 pH ≥7.60 & pCO2< 25: 0 pH ≥ 7.60 & pCO2< 40: 3 pH ≥ 7.60 & pCO2≥40: 12
Albumin (g/l)	<2.0: 11 2.0-2.4: 6 2.5-4.4: 0 ≥4.5: 4
Bilirubin (mg/dL)	<2.0: 0 2.0-2.9: 5 3.0-4.9: 6 5.0-7.9: 8 ≥8.0: 16

Table A10. APACHE Points for the Acute Physiology Score (continued)

APS Variable	**APS Points**
Blood Urea Nitrogen (mg/dL)	<17: 0 17-19: 2 20-39: 7 40-79: 11 ≥80: 12
Creatinine (mg/dL)	ARF if creatinine≥1.5 & urine<410 and dialysis=0 For ARF=1: <1.5: 0 ≥1.5: 10 For ARF=0: <0.5: 3 0.5-1.49: 0 1.5-1.94: 4 ≥1.95: 7
Glucose (mg/dL)	<40: 8 40-59: 9 60-199: 0 200-349: 3 ≥350: 5
Hematocrit	≤40: 3 41-49: 0 ≥50: 3
PaO2 (%) or Alveolar–arterial gradient* for intubated patients with $FiO_2 \geq 0.5$)	≤49: 15 50-69: 5 70-79: 2 ≥80: 0 If patient is intubated, and $FiO_2 \geq 0.5$, use Alveolar–arterial gradient <100: 0 100-249: 7 250-349: 9 350-499: 11 ≥500: 14
Sodium (mEq/L)	<120: 3 120-134: 2 135-154: 0 ≥155: 4

Table A10. APACHE Points for the Acute Physiology Score (continued)

APS Variable	**APS Points**
Urine Output (mL)	<400: 15 400-600: 8 600-899: 7 900-1499: 5 1500-1999: 4 2000-3999: 0 ≥4000: 1
WBC Count (x 10^9/L)	<1.0: 19 1.0-2.9: 5 3.0-19.9: 0 20.0- 24.9: 1 ≥25.0: 5
Glasgow Coma Scale	See Appendix 0A11 and Appendix Table A12.

Note: For intubated patients with $FiO_2 \geq 0.5$, we use Alveolar–arterial gradient (A-a gradient). The formula for A-a Gradient= $F_iO_2\left(P_{atm} - P_{H_2O}\right) - \frac{PCO_2}{R}$[133]. Here, F_iO_2 is the fraction of inspired oxygen, P_{atm} is the atmospheric pressure (760 mm HG at sea level), P_{H_2O} is the water partial pressure in alveolus ($= 47$ mmHg at sea level, 100% saturated), and R is the respiratory quotient (normally 0.8). To match the acute physiology score computed in the eICU database, we let R=1, although it is typically set to 0.8, and we assumed sea level[133].

Table A11. Glasgow Coma Score

Score	**Best eye response**
1	No eye opening
2	Eye opening to pain
3	Eye opening to verbal command
4	Eye opening spontaneously
Score	**Best verbal response**
1	No verbal response
2	Incomprehensible sounds
3	Inappropriate words
4	Confused
5	Oriented
Score	**Best motor response**
1	No motor response
2	Extension to pain
3	Flexion to pain
4	Withdrawal from pain
5	Localizing pain
6	Obeys command

Note: A Glasgow coma score of 13 or higher correlates with a mild brain injury, 9-12 is a moderate injury, and 8 or less a severe brain injury

Table A12. Acute Physiology Score Points for the Glasgow Coma Scale

Eyes (range: 1-4)	**Verbal (range: 1-5)**	**Motor (range: 1-6)**	**APS Score**
1	1	1 or 2	48
		3 or 4	33
		5 or 6	16
	2, 3, 4, or 5	1 or 2	29
		3 or 4	24
		5 or 6	Not clinically feasible
2, 3, or 4	1	1 or 2	29
		3 or 4	24
		5 or 6	15
	2 or 3	1 or 2	29
		3 or 4	24
		5	13
		6	10
	4	1, 2, 3, or 4	13
		5	8
		6	3
	5	1, 2, 3, 4, or 5	3
		6	0

APPENDIX B. R CODE

Simulations

```
library(plyr)
library(dplyr)
library(tidyverse)
library(parallel)
library(lme4)
library(lmerTest)
library(performance)
library("ProfileMatchit")

generate_data<-function(nt, nc, orig.SMD, orig.icc, treatment.effect){

#SET-UP THE PARAMETERS FOR THE DATA GENERATION
        #measured covariates based on chosen orig.SMD. options are either 0.1, 0.4, 0.3, 0.5
       if(orig.SMD==0.1){mu<-c(0.05, 0.10, 0.15, 0.05, 0.10, 0.15, 0.05, 0.10, 0.15, 0.10)} else
        if(orig.SMD==0.3){mu<-c(0.1, 0.2, 0.3, 0.4, 0.5, 0.1, 0.2, 0.3, 0.4, 0.5)} else
        if(orig.SMD==0.4){mu<-c(0.2, 0.3, 0.4, 0.5, 0.6, 0.2, 0.3, 0.4, 0.5, 0.6)} else
        if(orig.SMD==0.5){mu<-c(0.3, 0.4, 0.5, 0.6, 0.7, 0.3, 0.4, 0.5, 0.6, 0.7)} else
      {stop("ERROR: Choose another parameter for the covariate overlap (0.1, 0.3, 0.4, 0.5)")}

        #GENERATE THE DATA

        #Generate covariates for the treated group
        xt<-data.frame(x1=rnorm(nt, mean=mu[1], sd=1), x2=rnorm(nt, mean=mu[2], sd=1),
                x3=rnorm(nt, mean=mu[3], sd=1), x4=rnorm(nt, mean=mu[4], sd=1),
                x5=rnorm(nt, mean=mu[5], sd=1), x6=rnorm(nt, mean=mu[6], sd=1),
                x7=rnorm(nt, mean=mu[7], sd=1), x8=rnorm(nt, mean=mu[8], sd=1),
                x9=rnorm(nt, mean=mu[9], sd=1), x10=rnorm(nt, mean=mu[10], sd=1),
                x11=rnorm(nt, mean=0, sd=1),
                Treatment=1,
                eta=rnorm(nt, mean=0, sd=1))

        #Generate covariates for the control group
        xc<-data.frame(x1=rnorm(nc, mean=0, sd=1), x2=rnorm(nc, mean=0, sd=1),
                x3=rnorm(nc, mean=0, sd=1), x4=rnorm(nc, mean=0, sd=1),
                x5=rnorm(nc, mean=0, sd=1), x6=rnorm(nc, mean=0, sd=1),
                x7=rnorm(nc, mean=0, sd=1), x8=rnorm(nc, mean=0, sd=1),
                x9=rnorm(nc, mean=0, sd=1), x10=rnorm(nc, mean=0, sd=1),
                x11=rnorm(nc, mean=0, sd=1),
                Treatment=0,
                eta=rnorm(nc, mean=0, sd=1))
```

```
#Randomly distribute patients into 100 hospitals (treatment is all one hospital)
        xt$hospital.id=1
        xc$hospital.id=sample(2:100, nc, replace=TRUE)

        sample<-rbind(xt, xc)
          #parameters for each covariate in the outcome model
        beta<-c(1,1,1,1,1,1,1,1,1,1)

#Add a hospital level variance parameter from normal distribution. Variance set so that ICC is
desired level

hosp.var<-(sum(beta^2)+treatment.effect^2)*orig.icc/(1-orig.icc)

sample<-sample%>%
group_by(hospital.id)%>%
mutate(hosp.effect=rnorm(1, mean=0, sd=sqrt(hosp.var)))

#GENERATE THE OUTCOME
sample<-sample%>%
          mutate(Y=-1 + beta[1]*x1 + beta[2]*x2 + beta[3]*x3 + beta[4]*x4 + beta[5]*x5
               + beta[6]*x6 + beta[7]*x7 + beta[8]*x8 + beta[9]*x9 + beta[10]*x10
               + beta[11]*x11                    #unmeasured covariate
               + treatment.effect*Treatment      #Treatment effect
               + hosp.effect                     #variance associated with hospital
               + eta)                            #random error
        return(sample)
        }

       ## Matching

       runsim<-function(nt, nc, orig.SMD, orig.icc, nmatches, treatment.effect){

        r.square<-matrix(NA, ncol=1, nrow=(2*nmatches)+4)
        all.results<-data.frame()
        results<-data.frame()

        sample<-generate_data(nt, nc, orig.SMD, orig.icc, treatment.effect)

#Determine the amount of variation in outcome that can by explained by the measured covariates
#Output the adjusted R^2 for the measured covariates
r.square<-summary(lm(Y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+Treatment,
data=sample))$adj.r.squared

        mean<-aggregate(sample[,1:10], list(sample$Treatment), FUN=mean)
        treated<-sample[which(sample$Treatment==1), 1:10]
        sd.treated<-apply(treated,2, sd)
```

```
        all.smd<-abs(mean[1,-1]-mean[2,-1])/sd.treated
        all.smd<-apply(all.smd,1,mean)

#Estimate the treatment effect with multilevel model using all data, unadjusted for the measured
covariates
        model1<-lmer(Y~Treatment+(1|hospital.id), data=sample)
        estimate1<-coef(summary(model1))[2,1:2] #Treatment Estimate and SE
        estimate1[3]<-estimate1[1]-1.96*estimate1[2]
        estimate1[4]<-estimate1[1]+1.96*estimate1[2]
        estimate1<-as.data.frame(t(estimate1))

        #Output the ICC from the model
        icc<-icc(model1)[1]

        estimate1<-cbind(estimate1,icc)
        colnames(estimate1)<-c("Estimate", "SE", "lower", "upper", "icc")

        estimate1<-estimate1%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
              bias=Estimate-treatment.effect)

        estimate1$nmatches<-0
        estimate1$model<-"Multilevel Unadjusted"
        estimate1<-cbind(estimate1, all.smd)


#Estimate the treatment effect with multilevel model using all data, adjust for the measured
covariates
model1b<-lmer(Y~Treatment+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+(1|hospital.id),
data=sample)
        estimate1b<-coef(summary(model1b))[2,1:2] #Treatment Estimate and SE
        estimate1b[3]<-estimate1b[1]-1.96*estimate1b[2]
        estimate1b[4]<-estimate1b[1]+1.96*estimate1b[2]
        estimate1b<-as.data.frame(t(estimate1b))

        #Output the ICC from the model
         icc<-icc(model1b)[1]

        estimate1b<-cbind(estimate1b,icc)
        colnames(estimate1b)<-c("Estimate", "SE", "lower", "upper", "icc")

        estimate1b<-estimate1b%>%
        mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
              bias=Estimate-treatment.effect)

        estimate1b$nmatches<-0
```

```
        estimate1b$model<-"Multilevel Adjusted"
        estimate1b<-cbind(estimate1b, all.smd)

        #Model without random effect
        model2<-glm(Y~Treatment, data=sample)
        estimate2<-coef(summary(model2))[2,1:2] #Treatment Estimate and SE
        estimate2[3]<-estimate2[1]-1.96*estimate2[2]
        estimate2[4]<-estimate2[1]+1.96*estimate2[2]
        estimate2<-as.data.frame(t(estimate2))
        colnames(estimate2)<-c("Estimate", "SE", "lower", "upper")
        estimate2$icc<-NA

        estimate2<-estimate2%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                 bias=Estimate-treatment.effect)

        estimate2$nmatches<-0
        estimate2$model<-"Singlelevel Unadjusted"
        estimate2<-cbind(estimate2, all.smd)


        model2b<-glm(Y~Treatment+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10, data=sample)
        estimate2b<-coef(summary(model2b))[2,1:2] #Treatment Estimate and SE
        estimate2b[3]<-estimate2b[1]-1.96*estimate2b[2]
        estimate2b[4]<-estimate2b[1]+1.96*estimate2b[2]
        estimate2b<-as.data.frame(t(estimate2b))
        colnames(estimate2b)<-c("Estimate", "SE", "lower", "upper")
        estimate2b$icc<-NA

        estimate2b<-estimate2b%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                 bias=Estimate-treatment.effect)

        estimate2b$nmatches<-0
        estimate2b$model<-"Singlelevel Adjusted"
        estimate2b<-cbind(estimate2b, all.smd)

estimates<-rbind(estimate2, estimate2b, estimate1, estimate1b, make.row.names=FALSE)
estimates$controls<-nc

rm(estimate1, estimate2, estimate1b, estimate2b, model1b, model2b, model1, model2, icc,
all.smd)
```

```
#Cardinality Matching
#Loop through conducting one-to-one, 1-to-2, ..., 1-to-n cardinality matching.

targets<-apply(sample[,1:10],2,mean)
tols<-.05*sd.treated

for (L in 1:nmatches){
        m2<-ProfileMatchit::profilematchit(treat=sample$Treatment, covs=sample[,1:10],
                targets=targets, tols=tols, estimand = "ATT",
                ratio = L, #number of controls for each treated (1-to-L matching),
                solver="gurobi", time=5*60)

            #Output the matched sample
            matched2<-match.data(m2, data=sample)

            mean<-aggregate(matched2[,1:10], list(matched2$Treatment), FUN=mean)
            treated<-matched2[which(matched2$Treatment==1), 1:10]
            sd.treated<-apply(treated,2, sd)

            all.smd<-abs(mean[1,-1]-mean[2,-1])/sd.treated
            all.smd<-apply(all.smd, 1, mean)

            #Estimate the treatment effect with multilevel model
            model1<-lmer(Y~Treatment+(1|hospital.id), data=matched2)
            estimate1<-coef(summary(model1))[2,1:2] #Treatment Estimate and SE
            estimate1[3]<-estimate1[1]-1.96*estimate1[2]
            estimate1[4]<-estimate1[1]+1.96*estimate1[2]
            estimate1<-as.data.frame(t(estimate1))
            #Output the ICC from the model
            icc<-icc(model1)[1]

            estimate1<-cbind(estimate1,icc)
            colnames(estimate1)<-c("Estimate", "SE", "lower", "upper", "icc")

            estimate1<-estimate1%>%
              mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                   bias=Estimate-treatment.effect)

            estimate1$nmatches<-L
            estimate1$model<-"Multilevel Unadjusted"
            controls<-sum(matched2$Treatment==0)
            estimate1<-cbind(estimate1, all.smd, controls)
```

```
#Estimate the treatment effect with multilevel model, adjusted for covariates
model1b<-lmer(Y~Treatment+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+(1|hospital.id),
data=matched2)
        estimate1b<-coef(summary(model1b))[2,1:2] #Treatment Estimate and SE
        estimate1b[3]<-estimate1b[1]-1.96*estimate1b[2]
        estimate1b[4]<-estimate1b[1]+1.96*estimate1b[2]
        estimate1b<-as.data.frame(t(estimate1b))

        #Output the ICC from the model
        icc<-icc(model1b)[1]

        estimate1b<-cbind(estimate1b,icc)
        colnames(estimate1b)<-c("Estimate", "SE", "lower", "upper", "icc")

        estimate1b<-estimate1b%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                 bias=Estimate-treatment.effect)

        estimate1b$nmatches<-L
        estimate1b$model<-"Multilevel Adjusted"
        controls<-sum(matched2$Treatment==0)
        estimate1b<-cbind(estimate1b, all.smd, controls)

        #Model without random effect
        model2<-glm(Y~Treatment, data=matched2)
        estimate2<-coef(summary(model2))[2,1:2] #Treatment Estimate and SE
        estimate2[3]<-estimate2[1]-1.96*estimate2[2]
        estimate2[4]<-estimate2[1]+1.96*estimate2[2]
        estimate2<-as.data.frame(t(estimate2))
        colnames(estimate2)<-c("Estimate", "SE", "lower", "upper")
        estimate2$icc<-NA

        #Compute bias and coverage
        estimate2<-estimate2%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                 bias=Estimate-treatment.effect)

        estimate2$nmatches<-L
        estimate2$model<-"Singlelevel Unadjusted"
        estimate2<-cbind(estimate2, all.smd,controls)

        model2b<-glm(Y~Treatment+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10, data=matched2)
        estimate2b<-coef(summary(model2b))[2,1:2] #Treatment Estimate and SE
        estimate2b[3]<-estimate2b[1]-1.96*estimate2b[2]
        estimate2b[4]<-estimate2b[1]+1.96*estimate2b[2]
        estimate2b<-as.data.frame(t(estimate2b))
```

```
        colnames(estimate2b)<-c("Estimate", "SE", "lower", "upper")
        estimate2b$icc<-NA

        #Compute bias and coverage
        estimate2b<-estimate2b%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                 bias=Estimate-treatment.effect)

        estimate2b$nmatches<-L
        estimate2b$model<-"Singlelevel Adjusted"
        estimate2b<-cbind(estimate2b, all.smd,controls)

estimates<-rbind(estimates, estimate2, estimate2b, estimate1, estimate1b, row.names=NULL)
        results<-cbind(orig.SMD=orig.SMD, orig.icc=orig.icc, estimates, row.names = NULL)
        }

      #Largest possible subset

      m3<-ProfileMatchit::profilematchit(treat=sample$Treatment, covs=sample[,1:10],
                                         targets=targets, tols=tols,
                                         estimand = "ATT",
                                         ratio = Inf,
                                         solver="gurobi",
                                         time=5*60)

      #Output the matched sample
      matched3<-match.data(m3, data=sample)

      mean<-aggregate(matched3[,1:10], list(matched3$Treatment), FUN=mean)
      treated<-matched3[which(matched3$Treatment==1), 1:10]
      sd.treated<-apply(treated,2, sd)

      all.smd<-abs(mean[1,-1]-mean[2,-1])/sd.treated
      all.smd<-apply(all.smd, 1, mean)

      #Size of the Control Group
      controls<-sum(matched3$Treatment==0)

      #Estimate the treatment effect with multilevel model, unadjusted
      model1<-lmer(Y~Treatment+(1|hospital.id), data=matched3)
      estimate1<-coef(summary(model1))[2,1:2] #Treatment Estimate and SE
      estimate1[3]<-estimate1[1]-1.96*estimate1[2]
      estimate1[4]<-estimate1[1]+1.96*estimate1[2]
      estimate1<-as.data.frame(t(estimate1))
```

```
        #Output the ICC from the model
        icc<-icc(model1)[1]
          estimate1<-cbind(estimate1,icc)
        colnames(estimate1)<-c("Estimate", "SE", "lower", "upper", "icc")

        estimate1<-estimate1%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
              bias=Estimate-treatment.effect)

        estimate1$nmatches<-Inf
        estimate1$model<-"Multilevel Unadjusted"
        estimate1<-cbind(estimate1, all.smd, controls)

#Multilevel model, adjusted for covariates
model1b<-lmer(Y~Treatment+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+(1|hospital.id),
data=matched3)
        estimate1b<-coef(summary(model1b))[2,1:2] #Treatment Estimate and SE
        estimate1b[3]<-estimate1b[1]-1.96*estimate1b[2]
        estimate1b[4]<-estimate1b[1]+1.96*estimate1b[2]
        estimate1b<-as.data.frame(t(estimate1b))

        #Output the ICC from the model
        icc<-icc(model1b)[1]

        estimate1b<-cbind(estimate1b,icc)
        colnames(estimate1b)<-c("Estimate", "SE", "lower", "upper", "icc")

        estimate1b<-estimate1b%>%
          mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
              bias=Estimate-treatment.effect)

        estimate1b$nmatches<-Inf
        estimate1b$model<-"Multilevel Adjusted"
        estimate1b<-cbind(estimate1b, all.smd, controls)

        #Model without random effect, unadjusted
        model2<-glm(Y~Treatment, data=matched3)
        estimate2<-coef(summary(model2))[2,1:2] #Treatment Estimate and SE
        estimate2[3]<-estimate2[1]-1.96*estimate2[2]
        estimate2[4]<-estimate2[1]+1.96*estimate2[2]
        estimate2<-as.data.frame(t(estimate2))
        colnames(estimate2)<-c("Estimate", "SE", "lower", "upper")
        estimate2$icc<-NA
```

```
        #Compute bias and coverage
          estimate2<-estimate2%>%
            mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                   bias=Estimate-treatment.effect)
            estimate2$nmatches<-Inf
          estimate2$model<-"Singlelevel Unadjusted"
          estimate2<-cbind(estimate2, all.smd, controls)

          #Singlelevel model, adjusted for covariates
          model2b<-glm(Y~Treatment+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10, data=matched3)
          estimate2b<-coef(summary(model2b))[2,1:2] #Treatment Estimate and SE
          estimate2b[3]<-estimate2b[1]-1.96*estimate2b[2]
          estimate2b[4]<-estimate2b[1]+1.96*estimate2b[2]
          estimate2b<-as.data.frame(t(estimate2b))
          colnames(estimate2b)<-c("Estimate", "SE", "lower", "upper")
          estimate2b$icc<-NA

          #Compute bias and coverage
          estimate2b<-estimate2b%>%
            mutate(coverage=if_else(lower<treatment.effect & upper>treatment.effect, 1, 0),
                   bias=Estimate-treatment.effect)

          estimate2b$nmatches<-Inf
          estimate2b$model<-"Singlelevel Adjusted"
          estimate2b<-cbind(estimate2b, all.smd, controls)

          estimates<-rbind(estimate2,estimate2b, estimate1,estimate1b, row.names=NULL)
          resultsa<-cbind(orig.SMD=orig.SMD, orig.icc=orig.icc, estimates, row.names = NULL)
rm(estimate1, estimate2, estimate1b, estimate2b, model1, model2, icc, m3, matched3, all.smd,
controls)

all.results<-rbind(all.results, results, resultsa)
all.results<-cbind(all.results, r.square)
return(all.results)
}
```

eICU Data Cleaning

```
library(sqldf)
library (deSolve)
library(dplyr)
library(lubridate)
library(ggplot2)
library(plotly)
library(MLmetrics)
library(table1)
library(tidyverse)
library(purrr)
library(fastDummies)

patient = read("patient.csv.gz", show_col_types = FALSE) #200859
hospital = read("hospital.csv.gz", show_col_types = FALSE) #208
apachePatientResult = read("apachePatientResult.csv.gz", show_col_types = FALSE) #297064
apacheApsVar = read("apacheApsVar.csv.gz", show_col_types = FALSE) #171177
apachePredVar = read("apachePredVar.csv.gz", show_col_types = FALSE)#171177
#Exclude ICU readmissions
patient_filtered = patient %>%
filter(unitstaytype!="readmit") %>%
left_join(hospital, by="hospitalid")

#Categorize everyone 89 and older as 90 to protect patient privacy
patient_filtered = patient_filtered %>%
mutate(age_numeric = if_else(age == "> 89", "90", age) %>% as.numeric())

#Limit to those 16 and older
patient_filtered = patient_filtered %>%
filter(age_numeric >= 16)

#Add in APACHE variables
patient_filtered = patient_filtered %>%
left_join(apachePatientResult %>%
filter(apacheversion == "IVa" & apachescore>0) %>%
select(patientunitstayid, acutephysiologyscore, apachescore, predictedhospitalmortality,
                    actualhospitallos, actualhospitalmortality), by = "patientunitstayid")
#Limit to those without missing apache
patient_filtered = patient_filtered %>%
filter(!is.na(apachescore) & apachescore>0 & predictedhospitalmortality>0)

#Limit to patients with a diagnosis
patient_filtered = patient_filtered %>% filter(!is.na(apacheadmissiondx))
```

```
#APACHE score for age
patient_filtered = patient_filtered %>%
          mutate(age_score = case_when(age_numeric<=44 ~0,
                                       age_numeric>=45 & age_numeric<=59~5,
                                       age_numeric>=60 & age_numeric<=64~11,
                                       age_numeric>=65 & age_numeric<=69~13,
                                       age_numeric>=70 & age_numeric<=74~16,
                                       age_numeric>=75 & age_numeric<=84~17,
                                 age_numeric>=85~24))

#Chronic Conditions score
patient_filtered = patient_filtered %>%
          left_join(apachePredVar %>%
          select(patientunitstayid, admitdiagnosis, aids, hepaticfailure, lymphoma,
          metastaticcancer, leukemia, immunosuppression, cirrhosis, electivesurgery,
          thrombolytics, diedinhospital, diabetes), by = "patientunitstayid")

patient_filtered = patient_filtered %>%
           mutate(comorbid_score = case_when(
             electivesurgery==1 ~ 0L,
             aids==1 ~ 23L,
             hepaticfailure==1 ~ 16L,
             lymphoma==1 ~ 13L,
             metastaticcancer==1 ~ 11L,
             leukemia==1  ~ 10L,
             immunosuppression==1 ~ 10L,
             cirrhosis==1 ~ 4L,
             TRUE ~ 0L
  ))

#Acute Physiology Score
          #APACHE Scoring
          apacheApsVar2 = apacheApsVar %>%
            mutate(temp_score = case_when(
              temperature == -1 ~ 0L,
              temperature < 33.0 ~ 20L,
              temperature < 33.5 ~ 16L,
              temperature < 34.0 ~ 13L,
              temperature < 35.0 ~ 8L,
              temperature < 36.0 ~ 2L,
              temperature < 40.0 ~ 0L,
              temperature >= 40.0 ~ 4L,
              is.na(temperature) ~0L
             ))
```

```
apacheApsVar2 = apacheApsVar2 %>%
  mutate(map_score = case_when(
    meanbp == -1 ~ 0L,
    meanbp <   40 ~ 23L,
    meanbp <   60 ~ 15L,
    meanbp <   70 ~ 7L,
    meanbp <   80 ~ 6L,
    meanbp <  100 ~ 0L,
    meanbp <  120 ~ 4L,
    meanbp <  130 ~ 7L,
    meanbp <  140 ~ 9L,
    meanbp >= 140 ~ 10L,
    is.na(meanbp)~ 0L
  ))

apacheApsVar2 = apacheApsVar2 %>%
  mutate(hr_score = case_when(
    heartrate == -1 ~ 0L,
    heartrate < 40 ~ 8L,
    heartrate < 50 ~ 5L,
    heartrate < 100 ~ 0L,
    heartrate < 110 ~ 1L,
    heartrate < 120 ~ 5L,
    heartrate < 140 ~ 7L,
    heartrate < 155 ~ 13L,
    heartrate >= 155 ~ 17L,
    is.na(heartrate)~ 0L
  ))

apacheApsVar2 = apacheApsVar2 %>%
  mutate(rr_score = case_when(
    respiratoryrate == -1 ~ 0L,
    vent == 1 & respiratoryrate <=12 & respiratoryrate>=6 ~ 0L,
    respiratoryrate <   6 ~ 17L,
    respiratoryrate <  12 ~ 8L,
    respiratoryrate <  14 ~ 7L,
    respiratoryrate <  25 ~ 0L,
    respiratoryrate <  35 ~ 6L,
    respiratoryrate <  40 ~ 9L,
    respiratoryrate <  50 ~ 11L,
    respiratoryrate >= 50 ~ 18L,
    is.na(respiratoryrate)~ 0L
  ))
```

```
apacheApsVar2 = apacheApsVar2 %>%
  mutate(acidbase_score = case_when(
    ph == -1 |  pco2== -1~ 0L,
    ph < 7.20 & pco2 < 50 ~ 12L,
    ph < 7.20 ~ 4L,

    ph < 7.30 & pco2 < 30 ~ 9L,
    ph < 7.30 & pco2 < 40 ~ 6L,
    ph < 7.30 & pco2 <50 ~ 3L,
    ph < 7.30 & pco2 >= 50 ~ 2L,

    ph < 7.35 & pco2 < 30 ~ 9L,
    ph < 7.35 & pco2 < 45 ~ 0L,
    ph < 7.35 & pco2 >=45 ~ 1L,

    ph < 7.45 & pco2 <30 ~ 5L,
    ph < 7.45 & pco2 < 45 ~ 0L,
    ph < 7.45 & pco2 >= 45 ~ 1L,

    ph < 7.50 & pco2 < 30 ~ 5L,
    ph < 7.50 & pco2 < 35 ~ 0L,
    ph < 7.50 & pco2 < 45 ~ 2L,
    ph < 7.50 & pco2 >= 45 ~ 12L,

    ph < 7.60 & pco2 < 40 ~ 3L,
    ph < 7.60 & pco2 >=40 ~ 12L,

    ph >= 7.60 & pco2 < 25 ~ 0L,
    ph >= 7.60 & pco2 < 40 ~ 3L,
    ph >= 7.60 & pco2 >= 40 ~ 12L,
    is.na(ph)~ 0L
  ))

apacheApsVar2 = apacheApsVar2 %>%
  mutate(sodium_score = case_when(
    sodium == -1 ~ 0L,
    sodium <  120 ~ 3L,
    sodium <  135 ~ 2L,
    sodium <  155 ~ 0L,
    sodium >= 155 ~ 4L,
    is.na(sodium)~ 0L
  ))
```

```
        apacheApsVar2 = apacheApsVar2 %>%
        mutate(arf = if_else(creatinine >= 1.5 & urine < 410 & urine>-1 & dialysis==0, 1,0,
missing=0)) %>%
          mutate(creatinine_score = case_when(
            creatinine == -1 ~ 0L,
            arf == 1 & creatinine <  1.5 ~ 0L,
            arf == 1 & creatinine >= 1.5 ~ 10L,
            arf == 0 & creatinine <  0.5 ~ 3L,
            arf == 0 & creatinine <  1.5 ~ 0L,
            arf == 0 & creatinine <  1.95 ~ 4L,
            arf == 0 & creatinine >= 1.95 ~ 7L,
            is.na(creatinine)~ 0L
          ))

        apacheApsVar2 = apacheApsVar2 %>%
          mutate(uo_score = case_when(
            urine == -1 ~ 0L,
            urine < 400 ~ 15L,
            urine < 600 ~ 8L,
            urine < 900 ~ 7L,
            urine < 1500 ~ 5L,
            urine < 2000 ~ 4L,
            urine < 4000 ~ 0L,
            urine >= 4000 ~ 1L,
            is.na(urine)~ 0L
          ))

        apacheApsVar2 = apacheApsVar2 %>%
          mutate(bun_score = case_when(
            bun == -1 ~ 0L,
            bun <  17.0 ~ 0L,
            bun <  20.0 ~ 2L,
            bun <  40.0 ~ 7L,
            bun <  80.0 ~ 11L,
            bun >= 80.0 ~ 12L,
            is.na(bun)~ 0L
          ))

        apacheApsVar2 = apacheApsVar2 %>%
          mutate(hct_score = case_when(
            hematocrit == -1 ~ 0L,
            hematocrit < 41.0 ~ 3L,
            hematocrit < 50.0 ~ 0L,
            hematocrit >= 50.0 ~ 3L,
            is.na(hematocrit)~ 0L
          ))
```

```
apacheApsVar2 = apacheApsVar2 %>%
  mutate(wbc_score = case_when(
    wbc == -1 ~ 0L,
    wbc <   1.0 ~ 19L,
    wbc <   3.0 ~ 5L,
    wbc <  20.0 ~ 0L,
    wbc <  25.0 ~ 1L,
    wbc >= 25.0 ~ 5L,
    is.na(wbc)~ 0L
  ))

apacheApsVar2 = apacheApsVar2 %>%
  mutate(bilirubin_score = case_when(
    bilirubin == -1 ~ 0L,
    bilirubin >= 8 ~ 16L,
    bilirubin >= 5 ~ 8L,
    bilirubin >= 3 ~ 6L,
    bilirubin >= 2 ~ 5L,
    bilirubin < 2  ~ 0L,
    is.na(bilirubin)~ 0L
  ))

apacheApsVar2 = apacheApsVar2 %>%
  mutate(albumin_score = case_when(
    albumin == -1 ~ 0L,
    albumin <= 1.9 ~ 11L,
    albumin <= 2.4 ~ 6L,
    albumin <= 4.4 ~ 0L,
    albumin >= 4.5 ~ 4L,
    is.na(albumin)~ 0L
  ))

apacheApsVar2 = apacheApsVar2 %>%
  mutate(glucose_score = case_when(
    glucose == -1 ~ 0L,
    glucose < 40 ~ 8L,
    glucose < 60 ~ 9L,
    glucose < 200 ~ 0L,
    glucose < 350 ~ 3L,
    glucose >= 350 ~ 5L,
    is.na(glucose)~ 0L
  ))
```

```
        apacheApsVar2 = apacheApsVar2 %>%
          mutate(pao2_score = case_when(
            pao2 == -1 ~ 0L,
            pao2 < 50 ~ 15L,
            pao2 < 70 ~ 5L,
            pao2 < 80 ~ 2L,
            pao2 >=80 ~ 0L,
            is.na(pao2)~ 0L))

#If Fio2>50% use aa gradient

#Formula:
apacheApsVar2$fio2<-if_else(apacheApsVar2$intubated==0 & apacheApsVar2$fio2==-1, 21,
apacheApsVar2$fio2)

#atmospheric pressure (p.atm) is 760 at sea level
p.atm<-760

        apacheApsVar2 = apacheApsVar2 %>%
          mutate(aa_grade = (fio2/100)*(p.atm-47)-(pco2/1)-pao2)

        apacheApsVar2 = apacheApsVar2 %>%
          mutate(aa_grade_score = case_when(
            intubated==0 ~ 0L,
            intubated==1 & fio2>=50 & aa_grade < 100 ~ 0L,
            intubated==1 & fio2>=50 & aa_grade <250 ~ 7L,
            intubated==1 & fio2>=50 & aa_grade <350 ~ 9L,
            intubated==1 & fio2>=50 & aa_grade <500 ~ 11L,
            intubated==1 & fio2>=50 & aa_grade >=500 ~ 14L
          ))

        apacheApsVar2 = apacheApsVar2 %>%
          mutate(pao2_aa_score = case_when(
            intubated==0 ~ pao2_score,
            intubated==1 & fio2 < 50 ~ pao2_score,
            intubated==1 & fio2 >= 50 ~ aa_grade_score
          ))

        # APACHE scores for Glasgow Coma Scale
        apacheApsVar2 = apacheApsVar2 %>%
          mutate(gcs_score = case_when(
            eyes == -1 | verbal == -1 | motor == -1 ~ 0,
            eyes == 1 & verbal == 1 & motor <= 2 ~ 48,
            eyes == 1 & verbal == 1 & motor <= 4 ~ 33,
            eyes == 1 & verbal == 1 & motor <= 6 ~ 16,
            eyes == 1 & verbal > 1 & motor <= 2 ~ 29,
```

```
            eyes == 1 & verbal > 1 & motor <= 4 ~ 24,
            eyes == 1 & verbal > 1 & motor <= 6 ~ 99,

            eyes > 1 & verbal == 1 & motor <= 2 ~ 29,
            eyes > 1 & verbal == 1 & motor <= 4 ~ 24,
            eyes > 1 & verbal == 1 & motor <= 6 ~ 15,
            eyes > 1 & verbal <= 3 & motor <= 2 ~ 29,
            eyes > 1 & verbal <= 3 & motor <= 4 ~ 24,
            eyes > 1 & verbal <= 3 & motor == 5 ~ 13,
            eyes > 1 & verbal <= 3 & motor == 6 ~ 10,
            eyes > 1 & verbal == 4 & motor <= 4 ~ 13,
            eyes > 1 & verbal == 4 & motor == 5 ~ 8,
            eyes > 1 & verbal == 4 & motor == 6 ~ 3,

            eyes > 1 & verbal == 5 & motor <= 5 ~ 3,
            eyes > 1 & verbal == 5 & motor == 6 ~ 0,
            TRUE ~ 0
          ))

        #Glasgow Coma Score, rescaled
        apacheApsVar2 = apacheApsVar2 %>%
          mutate(Glasgow_rescaled=15-(eyes+verbal+motor))

        apacheApsVar2 = apacheApsVar2 %>%
          mutate(Glasgow_rescaled = if_else(
            eyes == -1 | verbal == -1 | motor == -1 , 15, Glasgow_rescaled
            ))

# Join to patient_filtered
eicu = patient_filtered %>%
  left_join(apacheApsVar2, by="patientunitstayid") #135987

#Diagnoses
dx<-read.csv("diagnoses categories.csv")

eicu<-eicu%>%
        left_join(dx %>%
                select(apacheadmissiondx, category, dx, postop, Diagnosis_code),
                by="apacheadmissiondx")

eicu$Diagnosis_code<-as.factor(eicu$Diagnosis_code)

#Create dummy variables for the categorical diagnosis
eicu <- dummy_cols(eicu, select_columns = 'dx')
```

```
#ED Admission
eicu$ED<-ifelse(eicu$hospitaladmitsource=="Emergency Department", 1, 0)

eicu$male<-if_else(eicu$gender=="Male", 1, 0)

eicu$gender[is.na(eicu$gender)]<-"Unknown"
eicu$ED[is.na(eicu$ED)]<-0

#Remove missing/other gender
eicu<-eicu[which(eicu$gender != "Unknown" & eicu$gender !="Other"),]
#Range of unit stays by hospital
eicu$hospid<-as.factor(eicu$hospitalid)

#Range of unit stays by hospital:
hosps<-eicu%>%
        group_by(hospitalid)%>%
        summarise(n=n())

#Limit to patients that went to hospitals with at least 300 hospitalizationshosps<-
hosps[which(hosps$n>=300), ]

eicu = hosps %>% left_join(eicu, by="hospitalid")
```

Analysis of eICU Data

```
devtools::install('C:/Users/mcgrathb/OneDrive - North Dakota University
System/NDSU/Dissertation/ProfileMatchit')

library("ProfileMatchit")
library(lme4)
library(table1)
library(dplyr)
library(gtsummary)
library(sqldf)
library(arm)
library(ggplot2)

#Read in the data and functions
eicu<-readRDS("eicu.rds")

#set the covariates
covars<-c("albumin_score", "bilirubin_score", "bun_score", "creatinine_score",
"glucose_score","hct_score", "acidbase_score", "sodium_score", "wbc_score",  "pao2_aa_score",
"uo_score", "hr_score", "map_score", "temp_score", "rr_score", "gcs_score", "male" , "postop",
"ED", "age_numeric", "comorbid_score","dx_CARDIOVASC", "dx_GI", "dx_HEMATO",
"dx_METABENDO",  "dx_MUSKELSKIN", "dx_NEUROLOGIC", "dx_RENALGENI",
"dx_RESPIRAT", "dx_SEPSIS", "dx_TRAUMA", "apachescore")

hosps<-eicu%>%
        group_by(hospitalid)%>%
        summarise(n=n())

#Create an indicator variable for hospital
hosps$hosp = seq.int(nrow(hosps))
```

One-to-One Matching

```
### one-to-one matching.

#Use each hospital as the profile. Find one match from the remaining control hospitals.
set.seed(1234)
datalist = list()

for (i in 1:nrow(hosps)) {
   eicu$case<-ifelse(eicu$hosp==i,1,0)

#Set the target to be the mean of the focal group (i.e., case hospital's population)
targets<-apply(eicu[which(eicu$case==1),covars], 2, mean)

#Let the tolerances be .05 standard deviations from the focal group
tols<-.05*apply(eicu[which(eicu$case==1),covars], 2, sd)

match.out<-ProfileMatchit::profilematchit(treat=eicu$case,
                                covs=eicu[,covars],
                                targets=targets,
                                tols=tols,
                                method = "cardinality",
                                estimand = "ATT",
                                ratio = 1,
                                verbose = FALSE,
                                solver="gurobi",
                                time=60*20)

#Output the matched sample
matched<-match.data(match.out, data=eicu)

matched$iteration<-i
        counts<-as.data.frame(table(matched$hosp))
        counts<-counts[which(counts$Var1 !=i),]
        average.cluster.size<-mean(counts$Freq)
        datalist[[i]]<-matched
}
allhosp1<-do.call(rbind, datalist)
allhosp1$iteration<-as.factor(allhosp1$iteration)

allhosp1$case <- factor(allhosp1$case, levels=c(0, 1), labels=c("Matched Controls", "Case
Hospital"))
allhosp1$diedinhospital <- factor(allhosp1$diedinhospital, levels=c(0, 1), labels=c("Alive",
"Expired"))
label(allhosp1$iteration) <- "Hospital"
saveRDS(allhosp1, "allhosp1.RDS")
```

```
##Compare in-hospital mortality rates at the case hospital relative to the comparison hospitals.
#Use a model that has a random intercept for hospital.

#Contains a random effect for matched hospital
modellist = list()

for (i in 1:nrow(hosps)) {

#Determine which hospitals were significantly higher than their benchmark.
  test<-allhosp1[which(allhosp1$iteration==i), ]

        model <- glmer(diedinhospital ~ case + comorbid_score + male + postop + ED +
        age_numeric +   dx_CARDIOVASC + dx_GI + dx_HEMATO + dx_METABENDO +
        dx_MUSKELSKIN +  dx_NEUROLOGIC + dx_RENALGENI +  dx_RESPIRAT +
        dx_SEPSIS + apachescore + (1 | hosp), data = test, family = binomial,  control =
        glmerControl(optimizer = "bobyqa", optCtrl=list(maxfun=1e6)), nAGQ =10)

  fixef<-as.data.frame(coef(summary(model)))
  fixef<-fixef[2,] #only keep the fixed effect associated with "case"

  fixef$sig<-ifelse(fixef$"Pr(>|z|)"<.05, 1,0)
  fixef$direction<-ifelse(fixef$"Estimate"<0, "lower","higher")
  fixef$iteration<-i
    modellist[[i]]<-fixef
}

allmodel1<-do.call(rbind, modellist)
saveRDS(allmodel1, "allmodel1.RDS")
table(sig=allmodel1$sig,allmodel1$direction)
```

Ten-to-One Matching

```
### ten-to-one Matching
#ten-to-one Mahalanbois distance matching was used in Silber, 2016

set.seed(1234)
datalist = list()

for (i in 1:nrow(hosps)) {
   eicu$case<-ifelse(eicu$hosp==i,1,0)

#Set the target to be the mean of the focal group (i.e., case hospital's population)
targets<-apply(eicu[which(eicu$case==1),covars], 2, mean)

#Let the tolerances be .05 standard deviations from the focal group
tols<-.05*apply(eicu[which(eicu$case==1),covars], 2, sd)

match.out<-ProfileMatchit::profilematchit(treat=eicu$case,
                                covs=eicu[,covars],
                                targets=targets,
                                tols=tols,
                                method = "cardinality",
                                estimand = "ATT",
                                ratio = 10,
                                verbose = FALSE,
                                solver="gurobi",
                                time=60*20)

#Output the matched sample
matched<-match.data(match.out, data=eicu)

matched$iteration<-i
        counts<-as.data.frame(table(matched$hosp))
        counts<-counts[which(counts$Var1 !=i),]
        average.cluster.size<-mean(counts$Freq)
        datalist[[i]]<-matched
}
allhosp10<-do.call(rbind, datalist)
allhosp10$iteration<-as.factor(allhosp10$iteration)

allhosp10$case <- factor(allhosp1$case, levels=c(0, 1), labels=c("Matched Controls", "Case
Hospital"))
allhosp1$diedinhospital <- factor(allhosp10$diedinhospital, levels=c(0, 1), labels=c("Alive",
"Expired"))
label(allhosp10$iteration) <- "Hospital"
saveRDS(allhosp10, " allhosp10.RDS")
```

```
##Compare in-hospital mortality rates at the case hospital relative to the comparison hospitals.
#Use a model that has a random intercept for hospital.

#Contains a random effect for matched hospital
modellist = list()

for (i in 1:nrow(hosps)) {

#Determine which hospitals were significantly higher than their benchmark.
  test<- allhosp10[which(allhosp10$iteration==i), ]

        model <- glmer(diedinhospital ~ case + comorbid_score + male + postop + ED +
        age_numeric +   dx_CARDIOVASC + dx_GI + dx_HEMATO + dx_METABENDO +
        dx_MUSKELSKIN +  dx_NEUROLOGIC + dx_RENALGENI +  dx_RESPIRAT +
        dx_SEPSIS + apachescore + (1 | hosp), data = test, family = binomial,  control =
        glmerControl(optimizer = "bobyqa", optCtrl=list(maxfun=1e6)), nAGQ =10)

  fixef<-as.data.frame(coef(summary(model)))
  fixef<-fixef[2,] #only keep the fixed effect associated with "case"

  fixef$sig<-ifelse(fixef$"Pr(>|z|)"<.05, 1,0)
  fixef$direction<-ifelse(fixef$"Estimate"<0, "lower","higher")
  fixef$iteration<-i
    modellist[[i]]<-fixef
}

allmodel10<-do.call(rbind, modellist)
saveRDS(allmodel10, " allmodel10.RDS")
table(sig= allmodel10$sig, allmodel10$direction)
```

Compute SMD of One-to-One and Ten-to-One Matched Samples

```
library(lme4)
library(arm)
library(ggplot2)
library(tidyverse)
library(smd)

allhosp1<-readRDS("allhosp1.RDS")
allhosp10<-readRDS("allhosp10.RDS")

#convert to factors

cols<- c("male" , "postop", "ED", "dx_CARDIOVASC", "dx_GI", "dx_HEMATO",
"dx_METABENDO", "dx_MUSKELSKIN", "dx_NEUROLOGIC", "dx_RENALGENI",
"dx_RESPIRAT", "dx_SEPSIS", "dx_TRAUMA")

allhosp1[cols] <- lapply(allhosp1[cols], factor)
allhosp10[cols] <- lapply(allhosp10[cols], factor)

keep<-c("iteration", "case", "albumin_score", "bilirubin_score", "bun_score", "creatinine_score",
"glucose_score","hct_score", "acidbase_score", "sodium_score", "wbc_score",  "pao2_aa_score",
"uo_score", "hr_score", "map_score", "temp_score", "rr_score", "gcs_score", "male" , "postop",
"ED", "age_numeric", "comorbid_score","dx_CARDIOVASC", "dx_GI", "dx_HEMATO",
"dx_METABENDO", "dx_MUSKELSKIN", "dx_NEUROLOGIC", "dx_RENALGENI",
"dx_RESPIRAT", "dx_SEPSIS", "dx_TRAUMA", "apachescore")


covars<-c("albumin_score", "bilirubin_score", "bun_score", "creatinine_score",
"glucose_score","hct_score", "acidbase_score", "sodium_score", "wbc_score",  "pao2_aa_score",
"uo_score", "hr_score", "map_score", "temp_score", "rr_score", "gcs_score", "male" , "postop",
"ED", "age_numeric", "comorbid_score","dx_CARDIOVASC", "dx_GI", "dx_HEMATO",
"dx_METABENDO", "dx_MUSKELSKIN", "dx_NEUROLOGIC", "dx_RENALGENI",
"dx_RESPIRAT", "dx_SEPSIS", "dx_TRAUMA", "apachescore")

meanSMD = list()

for (i in 1:113) {
    test<-allhosp1[which(allhosp1$iteration==i), keep ]
    #Compute the SMD for each variable used in the matching algorithm
    smd<-test %>%
      summarize_at(
        .vars = covars,
        .funs = list(smd = ~ smd(., g = case)$estimate))
```

```
meansmd<-as.data.frame(apply(abs(smd),1,mean))
  meansmd$iteration<-i
   meanSMD[[i]]<-meansmd
}

meanSMD<-do.call(rbind, meanSMD)
colnames(meanSMD)<-c("meanSMD", "iteration")
min(meanSMD[,1])
max(meanSMD[,1])
median(meanSMD[,1])

#SMD for the ten-to-one Matched cohort

meanSMD = list()
for (i in 1:113) {
    test<-allhosp10[which(allhosp10$iteration==i), keep ]
  #Compute the SMD for each variable used in the matching algorithm
  smd<-test %>%
    summarize_at(
      .vars = covars,
      .funs = list(smd = ~ smd(., g = case)$estimate))

  meansmd<-as.data.frame(apply(abs(smd),1,mean))
  meansmd$iteration<-i

  meanSMD[[i]]<-meansmd
}

meanSMD<-do.call(rbind, meanSMD)
colnames(meanSMD)<-c("meanSMD", "iteration")
min(meanSMD[,1])
max(meanSMD[,1])
median(meanSMD[,1])
```

Standard Regression Approach

```
library(lme4)
library(table1)
library(dplyr)
library(gtsummary)
library(sqldf)
library(arm)
library(ggplot2)
library(tidyverse)

n_percent = function(x, value = 1) {
  return(paste0(
    format(sum(x %in% value, na.rm = T), big.mark = ","), " (",
    format(round(sum(x %in% value, na.rm = T)/n()*100, digits = 1), nsmall = 1), "%)"))
}

median_iqr = function(x) {
  return(paste0(format(round(median(x, na.rm = T),2), nsmall = 0, big.mark = ","), " (",
                format(round(quantile(x, probs = .25, na.rm = T), 2), nsmall = 0), "-" ,
                format(round(quantile(x, probs = .75, na.rm = T),2), nsmall = 0), ")"))
}

median_range = function(x) {
  return(paste0(format(round(median(x, na.rm = T),2), nsmall = 0, big.mark = ","), " (",
                format(round(min(x, na.rm = T), 2), nsmall = 0), "-" ,
                format(round(max(x, na.rm = T),2), nsmall = 0), ")"))
}

eicu<-readRDS("eicu.rds")

covars<-c("albumin_score", "bilirubin_score", "bun_score", "creatinine_score", "glucose_score",
"hct_score"
          , "acidbase_score", "sodium_score", "wbc_score",  "pao2_aa_score", "uo_score"
          , "hr_score", "map_score", "temp_score", "rr_score", "gcs_score"
          , "male" , "postop", "ED", "age_numeric", "comorbid_score"
          ,"dx_CARDIOVASC", "dx_GI", "dx_HEMATO", "dx_METABENDO",
          "dx_MUSKELSKIN", "dx_NEUROLOGIC", "dx_RENALGENI", "dx_RESPIRAT",
          "dx_SEPSIS", "dx_TRAUMA",
          "apachescore", "acutephysiologyscore", "hosp")


data<-eicu[covars]
```

```
#Descriptives
table1(~age_numeric+factor(male)  + factor(postop)  + factor(ED)  + acidbase_score +
albumin_score + bilirubin_score + bun_score + creatinine_score + glucose_score + hct_score +
pao2_aa_score + sodium_score + uo_score + wbc_score + hr_score + map_score + rr_score +
temp_score + gcs_score + factor(dx_CARDIOVASC) + factor(dx_GI) + factor(dx_HEMATO) +
factor(dx_METABENDO) + factor(dx_MUSKELSKIN) + factor(dx_NEUROLOGIC) +
factor(dx_RENALGENI) + factor(dx_RESPIRAT) + factor(dx_SEPSIS) +
factor(dx_TRAUMA) + comorbid_score + acutephysiologyscore, data=data)

out<-data%>%
  group_by(hosp)%>%
  summarize(
    age_numeric=mean(age_numeric),
    male=mean(male),
    postop=mean(postop),
    ED=mean(ED),
    acidbase_score=mean(acidbase_score),
    albumin_score=mean(albumin_score),
    bilirubin_score=mean(bilirubin_score),
    bun_score=mean(bun_score),
    creatinine_score=mean(creatinine_score),
    glucose_score=mean(glucose_score),
    hct_score=mean(hct_score),
    pao2_aa_score=mean(pao2_aa_score),
    sodium_score=mean(sodium_score),
    uo_score=mean(uo_score),
    wbc_score=mean(wbc_score),
    hr_score=mean(hr_score),
    map_score=mean(map_score),
    rr_score=mean(rr_score),
    temp_score=mean(temp_score),
    gcs_score=mean(gcs_score),
    dx_CARDIOVASC=mean(dx_CARDIOVASC),
    dx_GI=mean(dx_GI),
    dx_HEMATO=mean(dx_HEMATO),
    dx_METABENDO=mean(dx_METABENDO),
    dx_MUSKELSKIN=mean(dx_MUSKELSKIN),
    dx_NEUROLOGIC=mean(dx_NEUROLOGIC),
    dx_RENALGENI=mean(dx_RENALGENI),
    dx_RESPIRAT=mean(dx_RESPIRAT),
    dx_SEPSIS=mean(dx_SEPSIS),
    dx_TRAUMA=mean(dx_TRAUMA),
    comorbid_score=mean(comorbid_score)
  )
```

```
hosp.medians<-out%>%
  summarise(
    age_numeric=median_iqr(age_numeric),
    male=median_iqr(male),
    postop=median_iqr(postop),
    ED=median_iqr(ED),
    acidbase_score=median_iqr(acidbase_score),
    albumin_score=median_iqr(albumin_score),
    bilirubin_score=median_iqr(bilirubin_score),
    bun_score=median_iqr(bun_score),
    creatinine_score=median_iqr(creatinine_score),
    glucose_score=median_iqr(glucose_score),
    hct_score=median_iqr(hct_score),
    pao2_aa_score=median_iqr(pao2_aa_score),
    sodium_score=median_iqr(sodium_score),
    uo_score=median_iqr(uo_score),
    wbc_score=median_iqr(wbc_score),
    hr_score=median_iqr(hr_score),
    map_score=median_iqr(map_score),
    rr_score=median_iqr(rr_score),
    temp_score=median_iqr(temp_score),
    gcs_score=median_iqr(gcs_score),
    dx_CARDIOVASC=median_iqr(dx_CARDIOVASC),
    dx_GI=median_iqr(dx_GI),
    dx_HEMATO=median_iqr(dx_HEMATO),
    dx_METABENDO=median_iqr(dx_METABENDO),
    dx_MUSKELSKIN=median_iqr(dx_MUSKELSKIN),
    dx_NEUROLOGIC=median_iqr(dx_NEUROLOGIC),
    dx_RENALGENI=median_iqr(dx_RENALGENI),
    dx_RESPIRAT=median_iqr(dx_RESPIRAT),
    dx_SEPSIS=median_iqr(dx_SEPSIS),
    dx_TRAUMA=median_iqr(dx_TRAUMA),
    comorbid_score=median_iqr(comorbid_score))%>%
  pivot_longer(cols=age_numeric:comorbid_score, names_to="variable")
```

```
###Multilevel model with full population###
full.model <- glmer(diedinhospital ~ albumin_score + bilirubin_score + bun_score +
creatinine_score + glucose_score + hct_score + acidbase_score + sodium_score + wbc_score +
pao2_aa_score + uo_score + hr_score + map_score + temp_score +  rr_score + gcs_score + male
+ postop + ED + age_numeric + dx_CARDIOVASC + dx_GI + dx_HEMATO +
dx_METABENDO + dx_MUSKELSKIN + dx_NEUROLOGIC + dx_RENALGENI +
dx_RESPIRAT + dx_SEPSIS  + comorbid_score
                + (1 | hosp), data = eicu,  family = binomial)
saveRDS(full.model, "full.model.rds")

summary(full.model)
fixed<-as.data.frame(coef(summary(full.model)))
RandomEffects <- as.data.frame(VarCorr(full.model))
ICC_between <- RandomEffects[1,4]/(RandomEffects[1,4]+pi^2/3)
#Note The residual deviance in logistic regression is fixed to (pi ^ 2) / 3
```

Creating the Figures

```
library(lme4)
library(arm)
library(ggplot2)
library(tidyverse)

#Read in the models from the standard approach, one-to-one matching, and ten-to-one matching
full.model<-readRDS("full.model.rds")
allmodel1<-readRDS("allmodel1.RDS")
allmodel10<-readRDS("allmodel10.RDS")

#Caterpillar plot
# Extract higher level residuals
ranef <-ranef(full.model)
ranef.se<-se.ranef(full.model)

ranef<-do.call(rbind, ranef)
ranef <- cbind(rownames(ranef), data.frame(ranef, row.names=NULL))

ranef.se<-do.call(rbind, ranef.se)

# Rank residuals
rank = rank(ranef[,2])

hi <- ranef[,2] + (1.96*ranef.se)
low <- ranef[,2]  - (1.96*ranef.se)

# Combine into data.frame
d <-data.frame(ranef,rank, hi, low)

d$sig<-if_else(0<d$X.Intercept..1 & 0>d$X.Intercept..2, 0, 1)
d$higher<-if_else(d$X.Intercept.>0, 1, 0)

table(significant=d$sig, higher=d$higher) #26 significantly higher, 20 significantly lower

#Compare the estimates from the one-to-one match and the standard regression approach.
d$iteration<-as.numeric(stringr::str_remove_all(d$rownames.ranef., "hosp."))

both<-d%>%left_join(allmodel1, by="iteration")
both$color<-if_else(both$sig.y==1 & both$direction=="lower", "#4daf4a",
                    if_else(both$sig.y==1 & both$direction=="higher", "#e41a1c", "black"))
#Significant by one-to-one matching

ggplot()+
geom_hline(yintercept=0,size=1, alpha=0.7,colour="gray", linetype="twodash")+
```

```
geom_pointrange(data=both,mapping=aes(x=rank, y=X.Intercept.,
ymin=X.Intercept..2,ymax=X.Intercept..1, color=color), ="identity", size=1)+
  theme_classic()+
  theme(axis.text.x=element_blank()) +
  theme(axis.text.y=element_text(size=12)) +
  theme(axis.title.y=element_text(size=12,vjust=1.5)) +
  theme(axis.title.x=element_text(size=12,vjust=-.5)) +
  theme(legend.position= "bottom") +
scale_x_continuous(name="Hospital Rank by Standard Regression Approach",
breaks=seq(1,113,1)) +
scale_y_continuous(name="Hospital Random Intercept") +
scale_color_manual(name="Result of one-to-one Matching",
labels=c("Significantly Lower", "Significantly Higher", "Not Significantly Different"),
values=c("#4daf4a","#e41a1c","black" ))+
# Plot margins and finally line annotations
theme(plot.margin = unit(c(1, 1, .5, .7), "cm"))
ggsave("caterpillar_1to1.tiff", dpi=300, width = 7, height = 5, units="in")

##Ten-to-One vs Standard Regression

both<-d%>%left_join(allmodel10, by="iteration")
both$color<-if_else(both$sig.y==1 & both$direction=="lower", "#4daf4a",
                    if_else(both$sig.y==1 & both$direction=="higher", "#e41a1c", "black"))
#Significant by ten-to-one matching

ggplot()+
geom_hline(yintercept=0,size=1, alpha=0.7,colour="gray", linetype="twodash")+
geom_pointrange(data=both,mapping=aes(x=rank, y=X.Intercept.,
ymin=X.Intercept..2,ymax=X.Intercept..1, color=color),  position="identity", size=1)+
  theme_classic()+
  theme(axis.text.x=element_blank()) +
  theme(axis.text.y=element_text(size=12)) +
  theme(axis.title.y=element_text(size=12,vjust=1.5)) +
  theme(axis.title.x=element_text(size=12,vjust=-.5)) +
  theme(legend.position= "bottom") +
scale_x_continuous(name="Hospital Rank by Standard Regression Approach",
breaks=seq(1,113,1)) +
scale_y_continuous(name="Hospital Random Intercept") +
scale_color_manual(name="Result of ten-to-one Matching", labels=c("Significantly Lower",
"Significantly Higher", "Not Significantly Different"), values=c("#4daf4a","#e41a1c","black"))+
# Plot margins and finally line annotations
theme(plot.margin = unit(c(1, 1, .5, .7), "cm"))
ggsave("caterpillar_10to1.tiff", dpi=300, width = 7, height = 5, units="in")
```

```
#Compare the estimates from the one-to-one match and the ten-to-one match.

both<-allmodel1%>%left_join(allmodel10, by="iteration")
both$color<-if_else(both$sig.x==1 & both$direction.x=="lower", "#4daf4a",
                    if_else(both$sig.x==1 & both$direction.x=="higher", "#e41a1c", "black"))
#Significant by one-to-one matching

both$shape<-if_else(both$sig.y==1 & both$direction.y=="lower", "A",
if_else(both$sig.y==1 & both$direction.y=="higher", "B", "C")) #Significant by one-to-one
matching

reg<-lm(formula = Estimate.y ~ Estimate.x, data=both)

#get intercept and slope value
coeff<-coefficients(reg)
intercept<-coeff[1]
slope<- coeff[2]

ggplot()+
  geom_point(data=both, aes(x=Estimate.x, y=Estimate.y, color=color, shape=shape),
             position="identity", size=2)+
  theme_classic()+
  theme(axis.text.y=element_text(size=12)) +
  theme(axis.title.y=element_text(size=12,vjust=1.5)) +
  theme(axis.title.x=element_text(size=12,vjust=-.5)) +
  theme(legend.position= "right") +
  scale_x_continuous(name="one-to-one Matching Estimate") +
  scale_y_continuous(name="ten-to-one Matching Estimate") +
  scale_color_manual(name="Result of one-to-one Matching",
labels=c("Significantly Lower", "Significantly Higher", "Not Significantly Different"),
values=c("#4daf4a","#e41a1c","black" ))+
  scale_shape_manual(name="Result of ten-to-one Matching",
labels=c("Significantly Lower", "Significantly Higher", "Not Significantly Different"),
values=c("triangle","square","circle"))+
  # Plot margins and finally line annotations
theme(legend.direction = "vertical", legend.box = "vertical")
ggsave("compare_1to1_10to1.tiff", dpi=300, width = 7, height = 5, units="in")
```