Stronger Instruments and Refined Covariate Balance in an Observational Study of the Effectiveness of Prompt Admission to the ICU in the UK*

Luke Keele[†] Steve Harris [‡] Samuel D. Pimentel[§] Richard Grieve[¶]

August 26, 2016

Abstract

Instrumental Variable (IV) methods can handle confounding due to differences in unobserved, as well as observed covariates, but in many settings, the IV is only weakly predictive of the treatment assignment. Near-far matching has been proposed for reducing bias when the IV is weak, by strengthening the instrument, and balancing observable characteristics. However, in settings with hierarchical data (e.g. patients nested within hospitals), or where several covariate interactions must be balanced, conventional near-far matching algorithms may fail to achieve the requisite covariate balance. We develop a new matching algorithm, that combines near-far matching with refined covariate balance, to balance large numbers of nominal covariates while also strengthening the IV. This extension of near-far matching is motivated by a UK case study that aims to identify the causal effect of receipt of prompt admission to the ICU on 90-day mortality.

*We thank Anirban Basu for comments and discussion.

[†]Associate Professor, McCourt School of Public Policy, Georgetown University, 37th & O St, NW Washington DC 20057 Email: ljk20@psu.edu, corresponding author.

[‡]UCL Hospital

[§]Department of Statistics, University of Pennsylvania, Philadelphia, PA, Email: spi@wharton.upenn.edu

[¶]Professor of Health Economics Methodology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, Email: richard.grieve@lshtm.ac.uk

1 Introduction

Clinical guidelines recommend that hospitals promptly admit critically-ill patients whose condition is deteriorating to the intensive care unit (ICU) to ensure closer monitoring and a higher level of acute care. However, in many hospitals, the ability to deliver ICU care is a constrained resource, since critical care is costly and there may be limited availability especially in a publicly funded health system like in the United Kingdom (Rhodes et al. 2012; Wunsch et al. 2014). Limited ICU capacity means that access to critical care in the UK is often constrained, and admission to the ICU for some patients may be delayed or refused.

Such delay or refusal may cause harm, but the effect of such harm is difficult to measure. A naive comparison of the mortality rate for those promptly admitted to the ICU versus patients that are delayed or refused admission would do little to answer the question. Patients admitted to the ICU may have higher mortality rates since they tend to be sicker than patients that are not admitted to the ICU. As such, ICU care may be associated with higher levels of mortality not because ICU care is ineffective but due to the patient population that tends to receive ICU care.

Ideally, the effect of prompt ICU admission would be settled using a randomized trial. However, evaluating the effect of delayed and refused admissions through a randomized clinical study is considered unethical because it is assumed that the costly, enhanced medical and nursing support provided in critical care is beneficial. Given the infeasibility of an randomized trial, the previous literature has attempted to establish the effectiveness of ICU care by using statistical methods to correct for confounding by indication (Simchen et al. 2007a; Robert et al. 2012). These studies have reported that prompt admission versus watchful-waiting leads to higher unadjusted mortality. However, the major concern in these studies is that we must assume that the assignment process into the ICU is fully observed. Given that around a third of patients admitted to critical care from general hospital wards are refused, there is a high likelihood that estimates of ICU effectiveness are subject to unmeasured confounding (hidden bias) (Harris et al. 2015).

In setting like this, one alternative is to find a natural experiment where circumstances create instances where some patients obtain prompt ICU care and others are denied treatment in a

haphazard manner. We exploit an instrument, which is a haphazard nudge or "encouragement" to treatment exposure (in this case prompt admission to the ICU). Specifically, we use the ICU occupancy-level at the time each patient was assessed for critical care as an instrument for the effect of prompt ICU admission (within 4 hours of assessment). We argue that the variation in ICU occupancy at the time of assessment tends to reflect local and temporal logistical barriers exogenous to the patients' prognosis. That is, while the bed availability in the ICU at the time of assessment will predict whether or not the patient is admitted promptly to the ICU, the corresponding occupancy level in the ICU will not have a direct effect on the patient's outcome. We also explore whether there are subgroup effects by baseline severity of illness.

We implement our study using a method know as near-far matching Baiocchi et al. (2010). This form of matching allows us to create a matched paired study design, where matched pairs are similar on observable characteristics but each pair is selected so that the individuals are distant according to the level of the IV. Hence units that are liable to weaken the instrument are discarded, analogous to applying exclusion criteria prior to randomization. However, in our study, hospitals form important blocks that may be correlated with unobserved confounders. To that end, we implement a near-far matching algorithm that also includes a method known as refined covariate balance, which allows us to account for these within hospital differences (Pimentel et al. 2015). As such, our study has novel methodological elements in that it is the first to combine these two forms of matching into a single algorithm.

The paper proceeds as follows. In Section 2, we outline the motivating example and the data. Section 3 provides an overview of the assumptions required for IV identification in the context of this example, and Section 4 provides details on near-far matching including the extension with refined covariate balance. Section 5 details the implementation, and Sections 6 and 7 the results. Finally, in Section 7, we conclude.

2 Motivating Example: Data and Summary Statistics

In UK hospitals, when patients on general hospitals wards are deemed to possibly be in need of critical care, floor staff call for a critical care assessment. At this time, staff from the ICU perform an assessment of whether the patient would improve under ICU care. If ICU care is thought to be beneficial that patient is admitted to the ICU, but not necessary transferred to the ICU. Depending on the bed space in the ICU, the patient may be admitted to critical care but there may be some delay until the patient is transferred and in some cases he or she may be admitted but never actually admitted. If a patient is not admitted and deteriorates further, he or she may be re-assessed and admitted at a later time.

Our study is based on a new data source: a prospective cohort study of deteriorating ward patients referred to critical care in 48 National Health Service (NHS) hospitals between 1 November 2010 and 31 December 2011). An advantage of this design compared to previous studies of the effectiveness of ICU care, is that we prospectively measure case-mix and outcomes for a cohort of patients who are assessed and judged suitable for prompt ICU admission, but only a subsample of these patients are actually admitted to the ICU. The data record the decision to admit a patient to ICU care and also the time from assessment to actual transfer into the ICU. The data also contain a critical care unit census, measuring ICU occupancy rates at the specific time of the patient's assessment. As we outline in more detail later, we use this measure of ICU bed availability at time of assessment at an instrumental variable. The primary endpoint was 7 and 28–day mortality.

As part of the data collection, a range of baseline covariates were also collected. These covariates were also recorded at the time of assessment and include age, septic diagnosis (0/1), and peri-arrest (0/1). Data were also collected on physiology measures. These measures include the Intensive Care National Audit & Research Centre (ICNARC) physiology score, the NHS National Early Warning Score (NEWS) which measures whether respiratory rate, oxygen saturations, temperature, systolic blood pressure, pulse rate, a level of consciousness vary from the norm, and the Sequential Organ Failure Assessment (SOFA) score which ranges from 0 to 24, with higher

scores indicating a greater degree of organ failure. The patient's existing level of care at assessment and recommended level of care after assessment were defined using the UK Critical Care Minimum Dataset (CCMDS) levels of care. These levels are 0 and 1 for normal ward care, 2 for care within a high dependency unit, 3 for care with intensive care unit. Finally, the data include three measures for periods when ICU capacity tends to be low. These measures include indicators for whether it was the weekend, out of hours—between 7 PM and 7 AM, and the months from November to February.

Our overall population is comprised of 15158 patients on general hospital wards that were assessed for admission to the ICU. That is, our study population consists of patients in a general hospital ward that were assessed for the possibility of critical care. Of these patients 2141 are excluded from the study due to the presence of a treatment limitation order which excluded the possibility of more aggressive treatment. For the remaining 13017 patients, six are excluded since data was missing on the availability of beds in the ICU at the time of assessment, which implies a study population of 13011 patients.

In our study, we define the treatment of interest as prompt admission to the ICU, which occurs if at the time of initial assessment he or she is admitted and to transferred to the ICU within four hours or less. This definition is in line with the definition of prompt admission in published guidelines in the UK (The Faculty of Intensive Care Medicine/The Intensive Care Society 2013). In our study population, 19.5% of the 13011 patients assessed were admitted promptly. Of the 10478 patients that were not promptly admitted 2432 or 23.2% were later transferred to the ICU. For those not promptly admitted to the ICU,mean time to the ICU is 22 hours and the median wait time is 10 hours. Finally, 245 patients (2.3%) were transferred within 5 hours, and 198 (1.9%) were transferred within 6 hours. Thus a small portion of patients receive ICU care just beyond the four hour window which defines our treatment. The critical care unit was full at the time of 1198 (8%) assessments. However, the ICU need not be completely full to discourage or delay admission. Patients that are in surgery are often directly admitted to critical care. As such, when a small number of ICU beds are available, some of these spaces may be reserved for

patients that will be directly admitted once surgery is completed.

3 Instrumental Variables and ICU Care

In a randomized encouragement design, some subjects are randomly encouraged to accept treatment, but some subset of the subjects fail to comply with the encouragement. Here the randomized encouragement to treatment exposure acts as the instrument. The method of IV, subject to a set of causal identification assumptions, provides an estimate of the causal effect of actual treatment exposure (Angrist et al. 1996). Analysts often seek to mimic the randomized encouragement design by finding instruments that occur in natural settings. Our study follows this template. We use ICU bed availability as an instrument for prompt admission to the ICU. That is, if few beds are available at the time of assessment, this should serve as a haphazard discouragement for prompt admission to the ICU. Any use of IV requires careful assessment of whether the identification assumptions are plausible in a specific context. Moreover, given that our instrument is not randomly assigned, as would be true in a randomized encouragement design, we must pay special attention to a possible interaction between assumptions. Therefore, we now turn to an examination of the IV assumptions in the context of our study.

3.1 Assessment of Identification Assumptions

For the number excess ICU beds to be a valid instrument, the five assumptions outlined by Angrist et al. (1996) must hold. These assumption are (1) no direct effect of instrument on outcome, also known as the exclusion restriction; (2) monotonicity; (3) the stable unit treatment value assumption (SUTVA); (4) the instrument must have a nonzero effect on the treatment and (5) ignorable (as-if random) instrument status.

We believe that occupancy levels meet the exclusion restriction for the following reasons. Firstly, the majority of patients studied are never admitted to critical care. Therefore, it is near impossible to see how the number of available beds in one specialist ward (the critical care unit)

could directly affect the outcome of a patient on second general ward who is never admitted. For those patients who are admitted, there is a concern that crowding in critical care during times of high occupancy might affect treatment. However, the instrument is the specific occupancy at the precise time of referral, and yet we know that occupancy both varies by time of day, and day of the week, and stochastically. Therefore subsequent crowding after admission is only weakly related to the occupancy at the time of referral. Moreover, the empirical evidence that does exist linking crowding directly to outcomes is conflicting (Gabler et al. 2013; Kahn et al. 2009).

To violate the monotonicity assumption it must the case that some patients are defiers. Defiers would be encouraged to receive prompt ICU care when there are few beds and discouraged to received prompt ICU care then there are many beds available. Such behavior by the admitting assessor would be fairly perverse. As such, we think it plausibly that there are few defiers. For SUTVA to hold, subjects outcomes must be unaffected by the assignment or treatment of other patients. While it is possible that one patient's outcomes are affected by whether another patient's treatment status, in most cases patients are not being assessed at the same time which reduces the likelihood of interference. Moreover two patients would have to be assessed at the same time in the same hospital since bed availability is a feature of a hospital. The instrument has a non-zero effect on prompt ICU admission. When we regressed an indicator for whether a patient was admitted to the ICU in four hours or less on the number of beds available, the t-statistic is 10.

We assess the final assumption by calculating covariate balance for patients above and below the median number of ICU beds available at the time of assessment. Note that this is only a partial assessment of this assumption since we cannot know whether unobservables are balanced. Table 2 contains means and standardized differences for baseline covariates. The standardized difference is the difference in means divided by the standard deviation before matching. Thus a standardized difference of one implies that the difference in means is equal to one standard deviation. We prefer standardized differences that are less than 0.20 and preferably 0.10 (Rosenbaum 2010). For many covariates, the imbalances are small, with just one standardized difference exceeding 0.10,

however, we do observe a number of differences between patients that are statistically significant.

Table 1: Balance Results For Units Above and Below the Median Number of ICU Beds Available at Time of Assessment.

	Less Than Median Beds[a] (N = 6114)	Greater Than Median Beds[a] (N=6897)	Std. Diff.	p-value
Age	64.95	65.40	-0.03	0.15
Male	0.52	0.53	-0.00	0.95
Sepsis 0/1	0.61	0.61	-0.01	0.67
Level of Care	1.05	1.07	-0.04	0.04
Rec'd Level of Care	1.37	1.44	-0.09	0.00
Peri-arrest 0/1	0.05	0.05	-0.03	0.12
Weekend	0.24	0.26	-0.05	0.01
Winter	0.32	0.20	0.28	0.00
Out of Hours	0.38	0.33	0.09	0.00
Icnarc Score	15.15	15.01	0.02	0.26
News Score	6.28	6.15	0.04	0.02
Sofa Score	3.18	3.12	0.03	0.14
Level of Care Missing	0.00	0.01	-0.08	0.00
Rec'd Level of Care Missing 0/1	0.00	0.01	-0.07	0.00

Note: [a]available at time of assessment for admittance to ICU. The measure of ICU bed availability had minimum value of zero and a maximum of 19, with a mean value of 4.3 and a median value of 4. Std. Diff. – standardized difference in means.

Given that patients are not perfectly balanced, our study is open to a threat from an interaction of IV assumptions. Small and Rosenbaum (2008) highlight that even in a study with a very large sample size, the problems created by a weak instrument worsen with small departures from assumption (5). In other words, when the instrument is weak, small departures from assumption (5) can produce very large biases *no matter how large the sample size*. Small and Rosenbaum (2008) also prove that a strong instrument is more robust to departures from ignorability even in smaller sample sizes. Thus they show that if assumption (5) does not hold, a smaller study with a stronger instrument will be less sensitive to bias than a weak instrument used in a much larger study. To avoid this threat we use a specific matching algorithm.

4 IV Matching with Refined Covariate Balance

Instruments in natural experiments are not ensured by the experimental design to be as-if randomly assigned. As such, analysts typically assume that the instrument is independent of unmeasured confounders conditional on measured covariates. Here, investigators must apply some method of statistical adjustment to any measured confounders. Typically analysts use two-stage least squares (2SLS) or some variant such as a "plug-in" estimator (Rivers and Vuong 1988; Nagelkerke et al. 2000; Palmer et al. 2008; Terza et al. 2008). Under this approach, covariate adjustment is straightforward as measured confounders are included in both the first and second stage regression models. Alternatively, covariate adjustment may proceed via matching. Covariate adjustment via matching has the advantage that we need not make arbitrary assumptions about the functional form of the relationship between covariates and outcomes. Such functional form assumptions can lead to biased estimates (Cai et al. 2011, 2012; Vansteelandt et al. 2011). This useful in our study, since both the exposure and outcome are binary covariates.

However, matching has one additional advantage when applied to IVs. Baiocchi et al. (2010) developed a novel form of matching as a design-based strategy for dealing with both covariate imbalance and weak instruments. Specifically, they applied a form of matching known as near/far matching to IV applications. Under most forms of matching, the goal is to find covariate pairs or subsets that are near on some measure of covariate distance such as the propensity score or Mahalanobis distance. Under a near/far match for IV designs, we retain the near matching typical of matched designs. That is, we seek to create pairs that have highly similar covariate values. In our application, we want a pair of patients that are very near each other in the covariate space such that they look identical but one received ICU care and the other did not. However, we would also prefer that one of these matched patients was strongly encouraged to receive ICU care since there were many beds available in the ICU, while the other patient was highly dissimilar in that few beds were available. Thus we would prefer to create pairs that have similar covariate values but dissimilar instrument values. In other words, we prefer pairs that are near on covariates but far apart on the instrument. The near/far matching algorithm in Baiocchi et al. (2010) creates

matched pairs that follows this template. This form of covariate adjustment directly addresses the result in Small and Rosenbaum (2008). It seeks to make pairs close such that assignment may appear as-if random in terms of observables, but the algorithm also seeks to make pairs increase discrepancy on the instrument to ensure that the instrument is not weak. We outline their method in more detail next. We extend their algorithm through the use of refined covariate balance constraints. The addition of refined covariate balance to the IV matching algorithm will allow us to control for important naturally occurring blocks in our application.

4.1 Notation

Under a near/far match, we use an algorithm that constructs pairs that seeks to maximize two objectives: near on covariates, far on the instrument. First, we introduce notation for the paired randomized encouragement design (Rosenbaum 1996, 2002). It is this experimental design that IV with matching mimics. This notation will help us to better convey the formal aspects of the algorithm. There are I matched pairs, $i = 1, \ldots, I$, and the units within matched pairs are denoted with $j \in \{1, 2\}$. We form these pairs by matching on observed covariates, $\mathbf{x}_{ij}$, which are measured before assignment to the instrument. Let W_{ij} denote the value of the instrument, excess beds in the ICU, for patient j in possible pair i. With a continuous instrument, such as excess beds, the ideal matched pair of subjects ik and il would have $\mathbf{x}_{ik} = \mathbf{x}_{il}$ but the difference, $W_{ik} - W_{il}$, will be large. That is, these units should be identical in terms of observed covariates but one of the units is strongly encouraged to take the treatment of ICU care while the other is not. Such a match creates comparable units with a large difference in terms of encouragement allowing for a stronger instrument.

4.2 Near/far Matching: A Review

Baiocchi et al. (2010) demonstrate how to use nonbipartite matching with penalties to implement the ideal IV match outlined above. However, alternative types of matching may be used. Zubizarreta et al. (2013) demonstrate how a near/far match may be implemented using integer

programming. Below we outline a third approach that relies on bipartite matching with penalties. As we outline below, our form of near/far matching allows for the use of refined covariate balance constraints which are critical to our application. One method for a near/far match relies on penalties. Penalties are used in matching to enforce compliance with a constraint whenever compliance is possible, and also to minimize the extent of deviation from a constraint whenever strict compliance is not possible. Under a near/far match, the matching algorithm attempts to minimize distances on observables within matched pairs subject to a penalty on instrument distance as measured by $W_{i1} - W_{i2}$, the distance between the observations in the matched pair on the instrument. The distance penalty, p, is defined as

$$p = \begin{cases} (W_{i1} - W_{i2})^2 \times c & \text{if } W_{i1} - W_{i2} < \Lambda \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where Λ is a threshold defined by the analyst. Note that the scale for Λ depends on the metric for W_{ij}. The penalty, p, is defined such that a smaller value of $W_{i1} - W_{i2}$ receives a larger penalty making those two units less likely to be matched, while c scales the penalty to that of the distance matrix. See Rosenbaum (2010, Sec. 8.4) for a discussion of penalties in matching. Matched distances on the instrument less than Λ receive larger penalties and thus are less likely to be matched. The result is that units that are alike on observables but more different on the instrument tend to be matched.

In most applications, however, increasing instrument distance tends to come at a cost to the near match on covariates. That is, as the algorithm forces pairs apart on the instrument, it becomes difficult to find balanced pairs. We remedy this problem through "sinks" (Lu et al. 2001). To improve balance, we use sinks to discard the observations that are hardest to match well. To eliminate e units that create the suboptimal matches, e sinks are added to the data before matching. We define each sink so that it has a zero distance between each unit and an infinite distance to all other sinks. This will create a distance matrix of size $(2I + e) \times (2I + e)$. The optimal nonbipartite matching algorithm pairs e units to the e sinks in such a way to minimize

the total distance between the remaining $I - e/2$ pairs. That is, by pairing a unit with a sink, the algorithm removes the e units that would form the e set of worst matches. Thus the optimal possible set of e units are removed from the matches.

As such, the analysts seeks a match where the penalty sufficiently strengthens the instrument, and the appropriate number of sinks yield acceptable within pair balance. Keele and Morgan (2016) show how to use weak instrument tests to guide the choice of the penalty and sinks. They use a grid search over both Λ and e. For each combination of Λ and e, the results from a weak instrument test are recorded. This allows the analyst to select the match which produces an instrument that is strong by the standards of weak instrument tests.

4.3 A Near/Far IV Match with Refined Covariate Balance

There are two additional features of our data that are directly relevant to the matching. First, we expect that there are important interactions between some of the covariates. Three of the covariates measure times when the ICU is expected to be more or less busy–during the winter, after hours, and on the weekend. While we seek to balance these covariate distributions within each matched pair, we might also seek to balance the interaction of these covariates as it is likely that ICU demand responds to the joint presence of these conditions.

Second, our study has important naturally occurring blocks in the form of hospitals. In National Health Service (NHS) hospitals, patients from general wards that are thought to be deteriorating are assessed for admission to the ICU. In many hospitals, the assessment for ICU admission is done by specially trained teams of experienced ICU nurses known as critical care outreach teams. In others, the assessment may simply be done by a single ICU physician or staff member. If the assessor believes the patient is in need of critical care, he or she will recommend that patient for ICU care. Depending on capacity in the ICU, the patient might be directly admitted to the ICU or there may be some delay. In short, assessment of whether a patient should be admitted to the ICU is a process that is focused in the hospital. In some locations, specially trained teams of nurses do this assessment, in other hospitals the assessment

is performed by the physician on duty in the ICU. As such, each hospital has a somewhat specific and perhaps idiosyncratic process that governs ICU admissions. Due to this fact, it may the be the case that important unmeasured covariates may be more similar within hospitals. Ideally, we would account for this variation by comparing a patient that was assessed for ICU admission at a time when the ICU had few beds available to another patient that was assessed with many beds available within the same hospital. Of course, we could simply exactly match to account for any confounders that are constant within hospital. The difficulty is that an exact match on hospital will tend to significant worsen balance on other covariates.

One alternative to an exact match is fine balance (Rosenbaum et al. 2007). Under fine balance, we require exact balance at all levels of the nominal variable but places no restriction on individual matched pairs-any one treated subject can be matched to any one control. When we use fine balance it means that the marginal distribution of a categorical covariate is exactly the same in treated and control groups. Fine balance constraints are not always feasible. A near fine balance constraint returns a finely balanced match when one is feasible, but minimizes the deviation from fine balance when fine balance is infeasible (Yang et al. 2012). In general, fine and near-fine balance are often used to balance a nominal variable with many levels, a rare binary variable or the interaction of several nominal variables.

In our study, most of the covariates are nominal and ideally we would seek to fine balance all these covariates. However, even with our large sample size we may encounter issues of sparsity. For example, we have 48 hospitals with 6 binary covariates and 2 nominal covariates with 4 levels, which implies 49,152 possible covariate levels. To deal with the sparsity that arises we use a method called refined covariate balance (Pimentel et al. 2015). Refined covariate balance is an extension of fine or near-fine balance designed to deal with large numbers of nominal covariates. Under refined covariate balance, we define a sequence of nested nominal covariates. The refined balance algorithm seeks to come as close to fine balance for the entire sequence of covariates by focusing on the joint interactions of the covariates. As such, if we add refined covariate balance to our match, we automatically ensure balance on the joint interaction of the nominal covariates.

For this project, we developed a new matching algorithm that combines near/far matching with refined covariate balance. We did this through the use of a reverse caliper. Caliper matching is a method that attempts to avoid poor matches by imposing a tolerance on the maximum distance between matched pairs (Cochran and Rubin 1973). For two subjects i and j, let P_i and P_j be a score on a distance metric such as the propensity score. Under a caliper, a match for subject i is selected only if $||P_i - P_j|| < \Lambda$, where Λ is a pre-specified tolerance. We can reverse the concept of a caliper and say a match for subject i is selected only if $||P_i - P_j|| > \Lambda$, where Λ remains a pre-specified tolerance. We use the reverse caliper in conjunction with matching to find units that are similar on observables, but we only keep a matched pair if it satisfies the reverse caliper such that the matched pair is very dissimilar in terms of the encouragement provided by the instrument. Thus we build a distance matrix applying a reverse caliper to the standard deviation of the instrument and penalizing matches that are too close on the instrument. We then match using the refined covariate balance algorithm.

The use of refined covariate balance does not change the basic tension between balance and instrument strength. As we strengthen the instrument, balance will tend to be worse. To increase instrument strength and balance covariates, we will again have to remove observations. We do so using optimal subset matching (Rosenbaum 2012). Optimal subset matching seeks to find the largest subset of treat for which the overall matched distance can be made smallest. To implement optimal subset matching we introduce a penalty parameter $\widetilde{\delta}$, which represents the cost of excluding a treated individual from the match. As the value for $\widetilde{\delta}$ increases, the match does not exclude anyone, and as as $\widetilde{\delta}$ is decreased, more and more units will be excluded. For a given value of $\widetilde{\delta}$ and fine balance constraints, the algorithm guarantees that the match produced has optimal refined balance among matches with the same number of units excluded. Thus the value $\widetilde{\delta}$ for serves a role identical to that of sinks. Thus we iterate over Λ and $\widetilde{\delta}$ to produce a match that has good balance and strengthens the instrument. Moreover, we use instrument weak tests to guide the selection of these two parameters.

5 The Match

5.1 How the matching was done

Next, we provide details on the matching. Prior to matching, we calculated the pairwise distances between the patients included in the sample. We used a rank-based Mahalanobis distance metric, which is robust to highly skewed variables (Rosenbaum 2010). For two covariates, a scale of the current level of care, and a scale for the recommended level of care a small fraction of data where missing. Instead of imputing these missing values based on a model, we use a method recommended by Rosenbaum (2010). To that end, we imputed missing values using the mean for that covariate. We then created a separate indicator for whether the value was missing. We then included the indicators for missing data in the match to ensure that missing values were balanced within matched pairs.

We first implemented a match that strengthened the instrument but did not include any refined covariate balance constraints. We conducted matches with several values for Λ and $\widetilde{\delta}$. Based on weak instrument tests, we settled on a match with $\Lambda = 1.5$ and $\widetilde{\delta} = 1000$. This implies that on average the discrepancy between matched pairs should be 1.5 standard deviations of the instrument. We later examine whether the study conclusions are sensitive to these choices. Next, we implemented a second match which retained the same parameters for strengthening the instrument but also included refined covariate balance constraints. We added refined covariate balance constraints for the hospital, the indicators for whether assessment occurred during the winter, after hours, or on the weekend, and the nominal measure for existing level of care at assessment and recommended level of care after assessment.

5.2 Balance Results

Table 2 contains means and standardized differences for both matches. The standardized difference on the instrument for both matches is approximately three. As such, both matches produce very similar results in terms of strengthening the instrument. The covariate balance is slightly

better in the second match that includes refined covariate balance. In this match, none of the standardized differences exceed 0.05, while in the first match one standard difference in 0.17 and two others are 0.10 or larger. These differences are a result of the refined covariate balance constraints which seek to balance the marginal distributions for these covariates. In general, levels of covariate balance are quite similar in terms of mean differences. Note that for both matches, we had to discard a large number of patients. The appendix contains a comparison between patients in each matched sample and the patients discarded by the match.

Table 2: Covariate balance and degree of encouragement in two matched comparisons. Std Diff = absolute standardized difference.

	Stronger Instrument W/o Refined Covariate Balance 4596 Pairs of Patients			Stronger Instrument With Refined Covariate Balance 2048 Pairs of Patients		
	Few Beds Available[a] Mean	More Beds Available[a] Mean	Std. Diff.	Few Beds Available[a] Mean	More Beds Available[a] Mean	Std. Diff.
Available ICU Beds	1.68	7.64	2.91	1.56	7.05	3.07
Age (years)	65.00	65.23	0.01	64.80	65.94	0.06
Male	0.53	0.53	0.01	0.54	0.54	0.00
Sepsis 0/1	0.63	0.62	0.00	0.62	0.62	0.00
CCMDS Level of Care - Level 0	0.13	0.11	0.07	0.10	0.10	0.01
CCMDS Level of Care - Level 1	0.69	0.71	0.04	0.70	0.70	0.01
CCMDS Level of Care - Level 2	0.17	0.16	0.02	0.18	0.18	0.01
CCMDS Level of Care - Level 3	0.01	0.01	0.05	0.01	0.01	0.00
Rec'd Level of Care - Level 0	0.08	0.04	0.17	0.04	0.04	0.01
Rec'd Level of Care - Level 1	0.55	0.55	0.00	0.53	0.52	0.02
Rec'd Level of Care - Level 2	0.28	0.30	0.05	0.32	0.32	0.01
Rec'd Level of Care - Level 3	0.08	0.09	0.04	0.11	0.11	0.02
Peri-arrest 0/1	0.04	0.05	0.07	0.05	0.05	0.02
Weekend 0/1	0.23	0.26	0.06	0.24	0.25	0.02
Winter 0/1	0.21	0.21	0.00	0.27	0.27	0.00
Out of Hours 0/1	0.36	0.34	0.05	0.36	0.34	0.05
Icnarc Score	15.23	15.07	0.02	15.59	15.57	0.00
News Score	6.28	6.18	0.03	6.42	6.28	0.05
Sofa Score	3.16	3.14	0.01	3.31	3.27	0.02
Level of Care Missing 0/1	0.00	0.01	0.10	0.00	0.00	0.04
Rec'd Level of Care Missing 0/1	0.00	0.01	0.11	0.00	0.00	0.01
Total Variation Distance		32.44			6.23	
Total Variation Distance Hospital Only		30.25			5.55	

Note: [a] at time of assessment for admittance to ICU. The measure of ICU bed availability had minimum value of zero and a maximum of 19, with a mean value of 4.3 and a median value of 4. Std. Diff. – standardized difference in means.

However, the comparison of means in Table 2 does little to convey the effect of applying the refined covariate balance constraints in the second match. In fact, we can be more formal about the how well the match balanced the marginal distributions of the nominal covariates on which we finely balanced in the second match. A nominal covariate k with L_k levels yields an $L_k \times 2$ contingency table with a column for treated observations and a column for control observations. One way to measure the discrepancy between the distributions on this covariate is to use the total variation distance (Pimentel et al. 2015). We denote β_{kl} for the difference in counts for row l of the table for covariate k, and $\sum_{l=1}^{L_k} |\beta_{kl}|$ is a measure of the difference between these two discrete probability distributions. We then express this as a proportion of the number of observations with each level of l to allow for comparisons across the two matches, since the sample sizes differ. This measure of total variation distance will be zero if there are no differences between the distribution, and will increase as the differences between the marginal distributions increases. While the total variation distance does not have a interpretable scale lower scores indicate smaller differences across the treated and control marginal distributions. We calculated the total variation distance twice. First, we calculated it for all the covariates that we finely balanced. Second, we calculated it justing using the marginal distributions for hospitals. These statistics are reported in the bottom panel of Table 2. In the match with refined covariate balance, the overall total variation distance is 6.23. In the second match, the total variation distance is 32.44, which is more than five times higher. We also observe that most of this imbalance stems from differences in the marginal distributions of hospitals. In both matches, the differences in the marginal distribution of hospitals accounts for 89% or more of the total variation distance. Table 3 contains the marginal distributions for hospital after the match based on refined covariate balance. While we were unable to finely balance this marginal distribution, in general, the discrepancy produced by near-fine balance is small.

6 Analyzing the Matched IV Design

6.1 Notation

After matching, there are I matched pairs for $i = 1, \ldots, I$, one whom is encouraged to have ICU care and the other which did not for $2I$ total units. We denote the patient with a lower value for W_{ij} as $Z_{ij} = 1$, and the other patient with a higher value for W_{ij} is denoted by $Z_{ij} = 0$, so that $Z_{i1} + Z_{i2} = 1$ for $i = 1, \ldots, I$. Let π_j denote the probability of being assigned to a value of the instrument for unit j. For two subjects, k and j matched so that observed covariates are similar, $\mathbf{x}_{ik} = \mathbf{x}_{ij}$, we assume that $\pi_j = \pi_k$. However, subjects may differ in the probability of treatment because they differ in terms of some unobserved covariate. That is, it may be the case that we failed to match on an important unobserved binary covariate u such that $\mathbf{x}_{ik} = \mathbf{x}_{ik}$, but possibly $u_{ik} \neq u_{ij}$.

Consistent with the potential outcomes framework (Neyman 1923; Rubin 1974), each patient has two potential responses under each level of Z_{ij}. That is, there are two responses denoted as (d_{Tij},d_{Cij}) and (r_{Tij},r_{Cij}), where the subscript T denotes treatment and C denotes control. Here, r_{Tij} and d_{Tij} are observed from the jth patient in pair i under $Z_{ij} = 1$, and r_{Cij} and d_{Cij} are observed from this same patient when $Z_{ij} = 0$. In our application, (r_{Tij},r_{Cij}) denotes mortality at 90 days, where 1 indicates dead and 0 indicates alive. The pair (d_{Tij},d_{Cij}) denotes prompt admission to the ICU if 1 and 0 if not. We do observe these potential outcomes, observed outcomes are reveal as $D_{ij} = Z_{ij}d_{Tij} + (1 - Z_{ij})d_{Cij}$ and $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$.

We test Fisher's sharp null hypothesis of no treatment effect on (r_{Tij},r_{Cij}). The sharp null asserts that $H_0 : r_{Tij} = r_{Cij}$ for all i and j, which implies that ICU care does not change the outcome for all patients. Moreover, the response will take the same value regardless of the value of Z_{ij}, which makes this model of effects consistent with the exclusion restriction. Informally, the exclusion restriction implies that instrument assignment Z_{ij} is related to the observed response $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$ only through the realized treatment D_{ij}. That is true here since $R_{ij} - \beta D_{ij}$ is a constant that does not vary with Z_{ij}.

6.2 The Generalized Effect Ratio

To estimate the effect of prompt ICU admission on mortality, we use the generalized effect ratio from Baiocchi et al. (2010). The generalized effect ratio is:

$$\lambda = \frac{\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij})}{\sum_{i=1}^{I}\sum_{j=1}^{2}(d_{Tij} - d_{Cij})} \tag{2}$$

which is the ratio of two average treatment effects. The average treated-minus-control difference provides unbiased estimates of the numerator and denominator. Under conditions such as $I \to \infty$ the ratio of these two unbiased effects is a consistent estimator for λ. The generalized effect ratio measures the relative magnitude of two treatment effects. The effect of excess beds on mortality and the effect of excess beds on whether a patient is promptly admitted for ICU care. Under the assumptions of monotonicity and the exclusion restriction, λ is the average decrease in mortality caused by ICU care among those patients who would receive ICU care only if there were plenty of beds available at the time of assessment (Angrist et al. 1996).

Next, we use the following terms derived in Baiocchi et al. (2010) to estimate an approximation to the randomization distribution for λ. The approximation is based on two terms, the first is:

$$T(\lambda_0) = \frac{1}{I}\sum_{i=1}^{I}\left\{\sum_{j=1}^{2} Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - \sum_{j=1}^{2}(1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij})\right\} = \frac{1}{I}\sum_{i=1}^{I} V_i(\lambda_0)$$

Next we define

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}\{V_i(\lambda_0) - T(\lambda_0)\}^2$$

Using these two terms, we can test $H_0 : \lambda = \lambda_0$ by comparing $T(\lambda_0)/S(\lambda_0)$ to a standard Normal cumulative distribution.

The generalized effect ratio is a form of randomization inference, where randomization forms

the "reasoned basis for inference" (Fisher 1935). One advantage of applying randomization inference to IV estimates becomes apparent with interval estimation in the presence of weak instruments (Rosenbaum 1999). An instrument is weak if d_{Tij} is close to or equal to d_{Cij} for most individuals ij. In other words, an instrument is weak instrument when most units ignore the encouragement to take the treatment. Under randomization inference, if the instrument is weak, the interval becomes longer and perhaps even infinite in length. In this case, a long confidence interval is a warning that the instrument provides little information about the treatment. As such, these confidence intervals provide clear warnings about the weakness of an instrument. Inference for the generalized effect ratio, is not exact, but is a large sample approximation to the exact randomization distribution. See Kang and Keele (2016) for a comparison between this method of inference and other more standard methods which rely on asymptotic approximations.

7 Results

Next, we report the estimated effect of prompt ICU care. We report outcome estimates for both matches in Table 4. For the first match that does not include refined covariate balance constraints, the point estimate for 7–day mortality is -0.03 with a 95% confidence interval of [-0.210, 0.144]. For 28–day mortality, the point estimate is -0.12, which is much larger, but the 95% confidence interval still includes zero, [-0.342, 0.088]. Next, we estimate outcomes for the match based on refined covariate balance. For 7–day mortality, the point estimate is -0.25 with a 95% confidence interval of [-0.642, 0.078]. For 28–day mortality, the point estimate is -0.19 with a 95% confidence interval of [-0.638, 0.216]. The point estimates from the second match are much larger than from the first match which did not include refined covariate balance constraints.

Both matches depend on our choices for the reverse caliper for instrument distance within matched pairs. As we noted above, we varied this parameter and selected a value based on balance and weak instrument tests. Next, we probe whether our outcome estimates our sensitive to this choice. To probe this possibility, we estimated outcome estimates for a match with the IV caliper

set to 1.0 and 2.0. These two matches bracket our final choice of 1.5 allowing us to observe if a somewhat weaker or stronger instrument match alter our results in any appreciable way. Note that for these alternative matches, we maintained the refined covariate balance constraints. For the match with a caliper of 1.0, the outcome estimates are 0.03 and 0.13 for 7– and 28-day mortality respectively. For the match with a caliper of 2.0, the outcome estimates are -0.02, and -0.21 for 7– and 28-day mortality respectively. Thus, we observe that the estimates are sensitive to the choice of the match parameters. However, in all cases the 95% confidence intervals span zero and are relatively wide. As such, precise comparisons of the different point estimates is not possible. We then repeated this exercise for the match which did not include the refined covariate balance constraints. For these matches, with the caliper set to 1.0, the outcome estimates are -0.09, and -0.14 for 7– and 28-day mortality respectively. When the caliper is set to 2.0, the outcome estimates are -0.04, and -0.09 for 7– and 28-day mortality respectively. Again, however, all 95% confidence intervals include zero, as such we are unable to reject the sharp null hypothesis. However, we generally find a positive benefit to ICU care which stands in contrast to the existing clinical literature, where ICU care tends to be associated with higher mortality rates.

8 Discussion

In this study, we examined the effect of prompt admission to a critical care unit on mortality. In our analysis, we took a design-based approach in that we implemented our statistical methods without reference to outcomes. Specifically, we implemented an instrumental variables analysis using matching methods. We used specialized matching algorithms to accomplish two tasks. First, we sought to find match pairs that were similar on covariates at baseline but dissimilar in terms of the instruments. Second, since the clinical process under investigation has hospital specific aspects, unobserved covariates may be more similar within hospitals. To that end, we used refined covariate balance to ensure that patients were near finely balanced on the hospital of admission. This is, to our knowledge, the first study to combine these two forms of matching.

Moreover, most of the effort in our study was spent on forming matched pairs. Once the matching is complete, we are able to use relatively simple methods to estimate treatment effects. The statistical methods, we use also account for binary outcomes without requiring strong functional form assumptions, and provide us with confidence intervals with appropriate coverage even when the instrument is weak.

This study has also advanced the literature on the effectiveness of ICU admission. The two largest extant studies (50322 patients (Chalfin et al. 2007) and 12268 patients (Simpson et al. 2005) were retrospective analyses of audit databases, and were unable to adjust for treatment selection bias. All but three studies were affected by exclusion and survival bias because there was no follow-up of patients assessed but not admitted. Among the three studies with complete follow-up, two focussed on populations wherein all, or nearly all, patients referred were admitted. One was a retrospective single centre study from the UK that reported a reduction in ventilator days, but no difference in survival for early admissions (O'Callaghan et al. 2012) The other was a prospective study of ten critical care units in France, that reported worse survival for the patients refused and re-referred, but not for all controls (Robert et al. 2012). One study with complete follow-up was performed in similar population to ours albeit with the addition of emergency department patients (Simchen et al. 2007b). These investigators evaluated outcomes for 749 deteriorating patients, and showed significantly better survival for ward patients admitted to critical care within 24 hours. This study suffered from the major concern that the results reflect that the patients were selected for critical care according to unmeasured prognostic factors.

Table 3: Marginal Distribution for Hospital Based on Refined Covariate Balance After Strengthening the Instrument

Control	Treated
41	41
27	27
82	82
89	89
63	83
49	49
90	43
10	20
60	60
91	89
4	4
10	11
3	12
16	0
35	35
30	31
74	74
18	18
136	136
2	8
17	14
66	66
53	90
23	23
39	39
54	54
42	48
11	11
1	7
10	18
119	119
67	67
13	13
77	77
58	58
9	10
79	79
64	47
17	8
60	54
103	104
28	16
28	34
38	38
10	10
13	13
19	19

Table 4: Estimated Effect of Prompt ICU Admission on 7 and 28 Day Mortality.

	Strong IV Only		
	Point Estimate	95% CI	p-value
7 Day Mortality	-0.031	[-0.210, 0.144]	0.73
28 Day Mortality	-0.122	[-0.342, 0.088]	0.256
	Strong IV & Refined Covariate Balancing		
	Point Estimate	95% CI	p-value
7 Day Mortality	-0.252	[-0.642, 0.078]	0.132
28 Day Mortality	-0.189	[-0.638, 0.216]	0.351

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010), "Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants," *Journal of the American Statistical Association*, 105, 1285–1296.

Cai, B., Hennessy, S., Flory, J. H., Sha, D., Have, T. R. T., and Small, D. S. (2012), "Simulation Study of Instrumental Variable Approaches With An Application to a Study of the Antidiabetic Effect of Bezafibrate," *Pharmacoepidemiology and Drug Safety*, 21, 114–120.

Cai, B., Small, D. S., and Have, T. R. T. (2011), "Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias," *Statistics in Medicine*, 30, 1809–1824.

Chalfin, D. B., Trzeciak, S., Likourezos, A., Baumann, B. M., Dellinger, R. P., study group, D.-E., et al. (2007), "Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit*," *Critical care medicine*, 35, 1477–1483.

Cochran, W. G. and Rubin, D. B. (1973), "Controlling Bias in Observational Studies," *Sankyha-Indian Journal of Statistics, Series A*, 35, 417–446.

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

Gabler, N., Ratcliffe, S., Wagner, J., Asch, D., Rubenfeld, G., Angus, D., and Halpern, S. (2013), "Mortality among patients admitted to strained intensive care units." *Am J Respir Crit Care Med*, 188, 800–806.

Harris, S., Singer, M., Rowan, K., and Sanderson, C. (2015), "Delay to admission to critical care and mortality among deteriorating ward patients in UK hospitals: a multicentre, prospective, observational cohort study," *The Lancet*, 385, S40.

Kahn, J., Ten Have, T., and Iwashyna, T. (2009), "The relationship between hospital volume and mortality in mechanical ventilation: an instrumental variable analysis." *Health Serv Res*, 44, 862–879.

Kang, H. and Keele, L. J. (2016), "A Comparison of Inferential Techniques for Instrumental Variables Methods," Unpublished Manuscript.

Keele, L. J. and Morgan, J. (2016), "How Strong is Strong Enough? Strengthening Instruments Through Matching and Weak Instrument Tests," *Annals of Applied Statistics*, Forthcoming.

Lu, B., Zutto, E., Hornik, R., and Rosenbaum, P. R. (2001), "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse," *Journal of the American Statistical Association*, 96, 1245–1253.

Nagelkerke, N., Fidler, V., Bersen, R., and Borgdorff, M. (2000), "Estimating treatment effects in randomized clinical trials in the presence of non-compliance," *Statistics in Medicine*, 19, 1849–1864.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

O'Callaghan, D. J., Jayia, P., Vaughan-Huxley, E., Gribbon, M., Templeton, M., Skipworth, J. R., and Gordon, A. C. (2012), "An observational study to determine the effect of delayed admission to the intensive care unit on patient outcome." *Crit Care*, 16, R173.

Palmer, T. M., Thompson, J. R., Tobin, M. D., Sheehan, N. A., and Burton, P. R. (2008), "Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses," *International Journal of Epidemiology*, 37, 1161–1168.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, Sparse Optimal Matching with Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Rhodes, A., Ferdinande, P., Flaatten, H., Guidet, B., Metnitz, P., and Moreno, R. (2012), "The variability of critical care bed numbers in Europe." *Intensive Care Med*, 38, 1647–1653.

Rivers, D. and Vuong, Q. H. (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.

Robert, R., Reignier, J., Tournoux-Facon, C., Boulain, T., Lesieur, O., Gissot, V., Souday, V., Hamrouni, M., Chapon, C., and Gouello, J. P. (2012), "Refusal of intensive care unit admission due to a full unit: impact on mortality." *American Journal of Respiratory and Critical Care Medicine*, 185, 1081–1087.

Rosenbaum, P. R. (1996), "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association*, 91, 465–468.

— (1999), "Using quantile averages in matched observational studies," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48, 63–78.

— (2002), "Covariance Adjustment In Randomized Experiments and Observational Studies," *Statistical Science*, 17, 286–387.

— (2010), *Design of Observational Studies*, New York: Springer-Verlag.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Mimimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatmetnt for Ovarian Cancer," *Journal of the American Statistical Association*, 102, 75–83.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 6, 688–701.

Simchen, E., Sprung, C. L., Galai, N., Zitser-Gurevich, Y., Bar-Lavi, Y., Levi, L., Zveibil, F., Mandel, M., Mnatzaganian, G., Goldschmidt, N., Ekka-Zohar, A., and Weiss-Salz, I. (2007a), "Survival of critically ill patients hospitalized in and out of intensive care." *Crit Care Med*, 35, 449–457.

— (2007b), "Survival of critically ill patients hospitalized in and out of intensive care." *Crit Care Med*, 35, 449–457.

Simpson, H. K., Clancy, M., Goldfrad, C., and Rowan, K. (2005), "Admissions to intensive care units from emergency departments: a descriptive study." *Emerg Med J*, 22, 423–428.

Small, D. and Rosenbaum, P. R. (2008), "War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases," *Journal of the American Statistical Association*, 103, 924–933.

Terza, J., Basu, A., and Rathouz, P. (2008), "Two-stage residual inclusion estimation: addressing endogeneity in health econometric Two-stage residual inclusion estimation: addressing endogeneity in health econometric Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling," *Journal of Health Economics*, 27, 531–543.

The Faculty of Intensive Care Medicine/The Intensive Care Society (2013), *Core Standards for Itensive Care*, London, UK: Intensive Care Society, 1st ed.

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. (2011), "On Instrumental Variables Estimaton of Causal Odds Ratios," *Statistical Science*, 26, 403–422.

Wunsch, H., Harrison, D., Jones, A., and Rowan, K. (2014), "The Impact of the Organization of High Dependency Care on Acute Hospital Mortality and Patient Flow for Critically Ill Patients." *Am J Respir Crit Care Med*.

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

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013), "Stronger Instruments via Interger Programming in an Observational Study of Late Preterm Birth Outcomes," *Annals of Applied Statistics*, 7, 25–50.

Appendices

A Appendix

Table 5: Balance Results for Observations Excluded from the Strong IV match

	Included in Match			Excluded From Match		
	Few Beds Available Mean	More Beds Available Mean	Std. Diff.	Few Beds Available Mean	More Beds Available Mean	Std. Diff.
Available ICU Beds	1.68	7.64	-2.91	1.94	4.25	-3.40
Age	65.00	65.23	-0.01	64.81	65.76	-0.05
Male	0.53	0.53	0.01	0.50	0.52	-0.04
Sepsis 0/1	0.63	0.62	0.00	0.56	0.59	-0.06
Level of Care	1.05	1.08	-0.05	1.03	1.05	-0.03
Rec'd Level of Care	1.37	1.45	-0.12	1.38	1.42	-0.05
Peri-arrest 0/1	0.04	0.05	-0.07	0.08	0.06	0.09
Weekend	0.23	0.26	-0.06	0.26	0.26	-0.01
Winter	0.21	0.21	0.00	0.67	0.19	1.10
Out of Hours	0.36	0.34	0.05	0.43	0.32	0.22
Icnarc Score	15.23	15.07	0.02	14.94	14.88	0.01
News Score	6.28	6.18	0.03	6.28	6.09	0.06
Sofa Score	3.16	3.14	0.01	3.23	3.09	0.06
Level of Care Missing	0.00	0.01	-0.10	0.01	0.01	-0.03
Rec'd Level of Care Missing 0/1	0.00	0.01	-0.10	0.01	0.01	0.02

Table 6: Balance Results for Observations Excluded from the Strong IV match

	Included in Match			Excluded From Match		
	Few Beds Available Mean	More Beds Available Mean	Std. Diff.	Few Beds Available Mean	More Beds Available Mean	Std. Diff.
Available ICU Beds	1.56	7.05	-3.07	1.84	6.28	-2.07
Age	64.80	65.94	-0.06	65.03	65.18	-0.01
Male	0.54	0.54	-0.00	0.52	0.52	-0.00
Sepsis 0/1	0.62	0.62	-0.00	0.60	0.61	-0.01
Level of Care	1.10	1.11	-0.00	1.02	1.05	-0.06
Rec'd Level of Care	1.49	1.51	-0.03	1.31	1.41	-0.14
Peri-arrest 0/1	0.05	0.05	-0.02	0.05	0.05	-0.03
Weekend	0.24	0.25	-0.02	0.23	0.26	-0.06
Winter	0.27	0.27	0.00	0.35	0.17	0.41
Out of Hours	0.36	0.34	0.05	0.38	0.33	0.11
Icnarc Score	15.59	15.57	0.00	14.93	14.77	0.02
News Score	6.42	6.28	0.05	6.20	6.10	0.03
Sofa Score	3.31	3.27	0.02	3.11	3.06	0.02
Level of Care Missing	0.00	0.00	-0.04	0.00	0.01	-0.09
Rec'd Level of Care Missing 0/1	0.00	0.00	-0.04	0.01	0.01	-0.08