RECSM

Research and Expertise Centre for Survey Methodology

Comparing different approaches for propensity score matching with clustered data: a simulation study

Bruno Arpino[1] and Massimo Cannas[2]

RECSM Working Paper Number 43

March 2015

http://www.upf.edu/survey/_pdf/RECSM_wp043.pdf

1 Department of Political and Social Sciences and Research and Expertise Centre for Survey Methodology (RECSM), Universitat Pompeu Fabra; Carrer Ramon Trias Fargas 25-27, 08005 Barcelona, Spain; bruno.arpino@upf.edu.

2 Department of Economic and Business Science, Università di Cagliari;
Viale Sant'Ignazio 84, Edificio Biblioteca di Economia, studio 14, 09123 Cagliari, Italy;
massimo.cannas@unica.it

Comparing different approaches for propensity score matching with clustered data: a simulation study

Bruno Arpino, *Universitat Pompeu Fabra*, bruno.arpino@upf.edu
Massimo Cannas, *University of Cagliari*, massimo.cannas@unica.it

Abstract. This article focuses on the implementation of propensity score matching for clustered data. Different approaches to reduce bias due to cluster-level confounders are considered and compared using Monte Carlo simulations. We investigated methods that exploit the clustered structure of data in two ways: in the estimation of the propensity score model (through the inclusion of fixed or random effects) or in the implementation of the matching algorithm. In addition to a pure within-cluster matching, we also assessed the performance of a "preferential" within-cluster matching. This approach first searches for control units to be matched to treated units within the same cluster. If matching is not possible within-cluster, then the algorithm searches in other clusters. All considered approaches successfully reduced the bias due to the omission of a cluster-level confounder. The preferential within-cluster matching approach, combining the advantages of within- and between-cluster matching, showed a relatively good performance both in the presence of big and small clusters and it was often the best method. An important advantage of this approach is that it reduces the number of unmatched units as compared to a pure within-cluster matching. We applied these methods to the estimation of the effect of caesarean section on the Apgar score using birth register data.

Keywords. propensity score; matching, clustered data, treatment effects, caesarean section, Apgar score.

1 Introduction

In observational studies, direct comparison of outcomes across treatment groups can give rise to biased estimates because groups being compared may be different due to lack of randomization. Subjects with certain characteristics may have higher probabilities than others to be exposed to the treatment. If these characteristics are also related to the outcome under investigation, an unadjusted comparison of the groups is likely to produce wrong conclusions about the treatment effect.

Propensity scores, defined as the probability to receive the treatment conditional on the set of observed variables, were introduced by Rosenbaum and Rubin [1] as a one-dimensional summary of the multidimensional set of covariates, such that when the propensity scores are balanced across the treatment and control groups, the distribution of all covariates are balanced across the two groups. In this way the problem of adjusting for a multivariate set of observed characteristics reduces to adjusting for the one-dimensional propensity score. (See Austin [2] for a review).

Importantly, propensity score methods can only ensure balance of background variables used in its estimation and consequently causal inferences based on these methods carry an assumption that no unobserved confounders exist. In medicine, as well as many other fields, data often show a clustered structure were individual units are nested into clusters (e.g., patients nested into hospitals as in our motivating case study). In these cases, bias can arise from omission in the propensity score model of individual and/or cluster-level confounders. Multilevel models for the outcome are commonly applied to study datasets with a clustered structure [3, 4]. However, in the context of unstructured data, it has been shown that, when treatment groups differ substantially, regression adjustment for the estimation of causal effects heavily rely on model extrapolation and can give rise to biased estimates [5].

Methods based on the propensity score (such as matching, weighting and stratification, see [6, 7] for a review) are a more robust alternative but only relatively recently has the treatment effect literature started to consider the implementation of propensity score methods with clustered data. Among the few works on the topic, Arpino and Mealli [8] showed, through an extensive simulation study, the benefit of using random- or fixed-effects models for the estimation of the propensity score to reduce the bias due to unmeasured cluster-level confounders when propensity scores are used to match treated and control individuals. Other studies have shown the benefits of exploiting the multilevel structure in the implementation of propensity score stratification [9] and weighting [10].

In this paper, we focus on propensity score matching and consider different approaches to take into account the clustered structure of the data with the aim of reducing the bias due to cluster-level confounders. We consider methods that exploit the information on the clusters to which units belong in two ways: in the estimation of the propensity score model *via* the inclusion of fixed or random effects; in the implementation of the matching algorithm.

As for the latter approach, the simplest idea to adjust for cluster effects is to implement the matching algorithm within each cluster. In fact, if we impose that treated and matched

control units must belong to the same cluster, we automatically achieve a perfect balance of all the observed and unobserved cluster-level characteristics. This approach was not considered in Arpino and Mealli [8] because their focus was on datasets characterized by very small-sized clusters, a case where within-cluster matching is not feasible.

When clusters sizes are big enough, within-cluster matching is a valid strategy but it can still imply the lost of many units that cannot find a match because the search is forced to be within clusters [11]. Discarding unmatched units is problematic because it may imply a change of the estimand [12]. In addition to a pure within-cluster matching, we also propose and assess the performance of an approach that has not been tested in previous studies. This approach first searches for control units to be matched to treated units within the same cluster. If matching is not possible within-cluster, then the algorithm searches in other clusters. This approach, that we define 'preferential' within-cluster matching, is expected to carry the benefits of pure within-cluster matching (in terms of bias reduction) and matching on the pooled dataset (in terms of minimizing the number of unmatched units).

We evaluate the performance of the different approaches in the context of a real case study and using a simulation set-up that mimics the available data. As a motivating case study we consider estimation of the effect of caesarian versus vaginal section on the probability of low Apgar score, a well-known measure of child's health. Our data are characterized by a strongly unbalanced structure with several hospitals with big sample sizes and some hospitals with small sample sizes. This type of data structure, that is not uncommon in real applications, has not been considered in previous Monte Carlo studies on propensity score techniques with clustered data. Moreover, our simulation experiments set variables distributions and effect sizes to mimic the distributions and associations observed in the empirical analysis. Thus, our approach, similarly to Huber et al. [13], is less prone to the standard critique of simulation studies that the chosen data generating processes are irrelevant for real applications. Given the lack of applications of propensity score methods with multilevel data, our paper serves also as an illustrative example on the usefulness of these methods in the context of a medical study.

In our motivating case study the relevant structure is represented by a hierarchy of two levels (individuals nested into hospitals) and we will consider this type of data structure in the following. However, the approaches we consider can be easily adapted to more complex structures. We focus on treatments assigned at the individual level. For propensity score methods applied to cluster randomized trials see Leyrat et al. [14, 15].

The rest of the article proceeds as follows. Section 2 introduces our motivating application and data. In section 3 we discuss alternative propensity score matching procedures for multilevel data. Section 4 presents a simulation study that mimics the data structure of the motivating example to assess the performance of the different methods under study. We also consider two variants of the baseline set-up. In section 5 we report results from the application of the proposed methods to the motivating case study. Section 6 offers some concluding remarks.

2 Motivating case study

As motivating case study, we consider the estimation of the causal effect of caesarean section (treatment) versus vaginal delivery (control) on the 5-minute Apgar score, a widely employed indicator of the clinical state of newborns [16, 17].

The 5-minute Apgar score is a clinical test performed on a newborn five minutes after birth. It is a composite measure of breathing effort, heart rate, muscle tone, reflexes, and skin color [18]. Each item is scored 0, 1, or 2, and thus the total score ranges from 0 to 10. Infants with a score higher than six are usually considered normal [19]. Low Apgar score is strongly associated with abnormal future development of the child [20] and infant mortality risk [21]. Following the literature, as outcome variable we consider a binary variable indicating whether the Apgar score is below the threshold of seven points ("low" Apgar score) or not. In our sample the proportion of low Apgar scores is 6.75 ‰. This is an example of low-incidence binary outcome for which propensity score matching has been found to perform considerably better than logistic regression adjustment on unstructured data [22].

The treatment considered in this study is the caesarean section, arguably one of the most common obstetric procedures. The intervention rate of caesarean delivery has grown substantially in the last decades [23]. Nonetheless, there is concern about whether the current high rates of caesarean sections are justified because the procedure is not without risk and previous studies reported negative effects on both the mother and the infant [24, 25]. In particular, some studies also found a negative effects of caesarean sections on the Apgar score [26].

Similarly to several medical studies, in our empirical analyses we use data collected in different sites (hospitals in our case). The dataset we consider contains information on deliveries occurred in the 22 hospitals of the Italian region of Sardinia in 2010 and 2011. The source is the official abstract of the birth event which is filled by physicians after each birth. This form is divided in three parts containing information on the mother, the pregnancy and the infant. Our dataset includes all hospitalized births in the period 2010-2011.

Following other observational studies on caesarean section (e.g., [23, 27]), to better isolate the effect of caesarian section on the target variable, we focus on the subset of non-complicated pregnancies. In particular, we selected nulliparous women at 32 or more weeks of gestational age with a singleton and living infant in vertex (head-down) position and without birth anomalies. We further restrict the sample to mothers aged between 15 and 44. These selection criteria are part of the matching strategy in the sense that they avoid including in the working sample women with extremely high probabilities of undergo a caesarean section and that could not find comparable control units. Our working sample (see Table 1) contains 14,757 observations clustered in twenty hospitals (two hospitals were removed since after applying the selection criteria they only contained treated or untreated women).

A simple unadjusted comparison of Apgar scores by mode of delivery indicates a negative effect of caesarean section. In fact, in our sample the prevalence of low Apgar scores

is 10.9 ‰ and 5.2 ‰ among treated and control women, respectively. Obviously, this unadjusted comparison ignores differences in individual and hospital characteristics between the groups of mothers that undergo caesarean or vaginal delivery. Indeed, previous studies have identified several clinical and socio-economic variables that proved to be associated with caesarean section and Apgar scores. Additional to individual-level confounders, the literature suggested the relevance of hospital-level factors both on the decision of taking a medical treatment and on medical outcomes. For example, Caceras et al. [27] and Bragg et al. [23] found that even after adjusting for socio-demographic and clinical factors, the rates of cesarean deliveries varied according to the hospital, suggesting the importance of hospital practices and culture in determining a hospital cesarean rate. Examples of relevant hospital-level variables include obstetricians practices, physicians preferences and guidelines promoting or restricting the liberal use of caesarean sections. Substantial variability in the outcomes of medical treatments across hospitals after controlling for clinical variables have been found due to differences in the quality of care [28, 29].

In our dataset, overall, 35% of births were caesarean, with a range of 11% to 64% across hospitals (see Table 1). Substantive variation across hospitals was also observed in the prevalence of low Apgar scores, which ranged from 0 to 16.5 cases every 1,000 deliveries, with an average of 6.75. Apart from an uneven distribution of individual-level risk factors, these variations across hospitals may indicate the presence of hospital-level important factors. In general it is difficult or impossible to measure all hospital characteristics that can impact on the probability to receive a treatment and on its outcomes, as they are not available in official forms. Our case is no exception and thus standard propensity score matching would fail to adjust for hospital-level omitted variables. In the next section we discuss strategies that exploit the information of the hospital where each mother has delivered to adjust for unobserved hospital-level confounders.

3 Propensity score matching with clustered data

Consider a two-level data structure where N individual-level units, indexed by i ($i = 1, 2, ..., n_j$), are nested in J second-level units (clusters), indexed by j ($j = 1, 2, ..., J$). We consider a binary treatment administered at the individual level, T, and an outcome variable, Y also measured at the individual level. Confounders can be first (X) or second-level (Z) variables.

Under the potential outcome framework, let $Y_{ij}(t)$ be the potential outcome if unit ij was assigned to treatment t, $t \in \{0, 1\}$. The fact that potential outcomes are labelled only by ij and t corresponds to the Stable Unit Treatment Value Assumption (SUTVA, [30]). This usually-invoked assumption requires that potential outcomes for a unit are not affected by the treatment received by other units, and there are no hidden versions of the treatment. Under SUTVA, for each unit ij, an individual causal effect is defined as a comparison of $Y_{ij}(1)$ with $Y_{ij}(0)$, yet only one of the two potential outcomes is observed depending on the value taken by T_{ij}.

Usually, the Average Treatment effect on the Treated (ATT) is considered as an in-

teresting summary of individual causal effects: $ATT = E(Y_{ij}(1) - Y_{ij}(0) \,|\, T_{ij} = 1)$. To identify the ATT with observational data, the following assumptions are often invoked:

- Unconfoundedness: $Y(1), Y(0) \perp T|(X, Z)$;

- Overlap: $0 < P(T = 1 \,|\, (X, Z)) < 1$.

Unconfoundedness asserts that the probability of assignment to a treatment does not depend on the potential outcomes conditional on observed covariates [1]. Unconfoundedness essentially assumes that within subpopulations defined by values of the covariates, we have random assignment of the treatment. This assumption is often referred to also as selection on observables because it rules out the role of unobserved variables [6]. Overlap implies that, for all possible values of the covariates there is a positive probability of receiving each treatment. Usually, analyses are restricted to the common support of covariates across treatment groups, where this assumption is met.

Under the previous assumptions, adjustment on the propensity score is sufficient to eliminate bias due to observed confounders [1]. The propensity score, e, is defined for each unit as the probability to receive the treatment given its covariate values: $e_{ij} = P(T_{ij} = 1 \,|\, (X_{ij}, Z_j))$. Rosenbaum and Rubin [1] proved that the propensity score is a balancing score, i.e., $(X, Z) \perp T \,|\, e(X, Z)$, meaning that at each value of the propensity score the distribution of the covariates defining the propensity score is the same in the treated and control groups. They also showed that if unconfoundedness holds conditioning on covariates it also holds conditioning on the propensity score, i.e., $Y(1), Y(0) \perp T \,|\, e(X, Z)$. These results justify adjustment on the propensity score instead of on the full multivariate set of covariates.

Usually, in observational studies the propensity score is not known and must be estimated from the data. Parametric models, such as logit or probit models, with inclusion of interactions and higher order terms are commonly used [31]. An incorrectly estimated propensity score may fail to respect the balancing property [5]. Our focus is not on misspecification of the functional form of the propensity score model but on the bias caused by omitted cluster-level confounders. If one or more variables affecting the selection into treatment and potential outcomes are not observed, then unconfoundedness is violated and ATT estimators based on the propensity score will be biased. In the following we shall assume that we have good measurement on all individual-level confounders, X, but we have no information on cluster-level confounders. For simplicity, we consider that all cluster-level effects are summarized by one cluster-level variable, Z, which is unobserved. The methods we discuss in the following can be adapted when some observed cluster-level variables are observed and others are not.

Among propensity score methods available to adjust for an unbalanced distribution of covariates between treated and control groups, we consider propensity score matching (PSM). In particular, we consider one-to-one nearest neighbor matching within a maximum distance (caliper) of 0.20 standard deviations of the estimated propensity score [32]. For each treated unit in the sample, the algorithm searches for the closest control unit

in terms of propensity score. If no control unit is available in the range defined by the caliper, the treated unit is discarded from the working sample. We considered matching with replacement, where the same control unit can be used several times as a match. Matching with replacement is expected to improve the quality of matches and therefore to reduce bias [33]. However, a bias-variance trade-off emerges because matching with replacement increases variance of estimates [34]. Since our main focus is on the bias of the estimators we considered matching with replacement.

When the dataset has a 2-level structure one can consider different ways of implementing PSM. The methods we compare are as follows:

A) Single-level propensity score model; matching on the pooled dataset;

B) Single-level propensity score model; within-cluster matching;

C) Single-level propensity score model; preferential within-cluster matching;

D) Random-effects propensity score model; matching on the pooled dataset;

E) Fixed-effects propensity score model; matching on the pooled dataset.

In approach A, we use a single-level logit model to estimate the propensity score:

$$logit(e_{ij}) = \alpha_0 + X_{ij}\beta. \tag{1}$$

Then, a one-to-one caliper matching on the estimated propensity scores is implemented on the pooled dataset. Formally, let I_1 and I_0 denote the set of treated and control units, respectively, and let indicate with A_{rj} the set of control units matched to $rj \in I_1$ (note that ij indicates a generic unit in cluster j while rj indicates a generic treated unit in cluster j):

$$A_{rj} = \left\{ kj' \in I_0 : \hat{e}_{kj'} = \min_{kj' \in I_0} |\hat{e}_{rj} - \hat{e}_{kj'}| < 0.20\hat{\sigma}_e \right\}. \tag{2}$$

After (2) has been constructed for all units in the treatment group, the matched dataset M is built:

$$M = \{rj : A_{rj} \neq \emptyset\} \cup \left\{ \bigcup_{rj} A_{rj} \right\} \tag{3}$$

that is, all treated units that successfully found a matched control unit and all matched control units form the matched dataset, M, while all other units are discarded. The ATT estimator takes the form:

$$\widehat{ATT} = \frac{1}{card(M)} \left\{ \sum_{rj \in I_1 \cap M} \left(Y_{rj} - \sum_{kj' \in I_0} Y_{kj'} w(rj, kj') \right) \right\}, \tag{4}$$

where $w(rj, kj')$ is the weight assigned to the control unit kj' in the estimation of the unobserved potential outcome, $Y(0)$, for the treated unit rj. In principle, in a nearest neighbour matching, at most one control unit is matched to each treated unit. However, it is possible that several control units have the same value of the propensity score within the caliper ("ties"). In this case A_{rj} will contains several units and each control unit in A_{rj} enters the (4) with a weight equal to $1/card(A_{rj})$. Control units that are not matched to rj are assigned a weight of zero in (4).

Model (1) is a simple single-level logit model that ignores the multilevel data structure and in fact the subscript ij is used only for notational consistency with the following methods. The multilevel structure is also ignored in the implementation of the matching (2). In fact in the creation of the matched pairs the cluster membership is ignored; there is no constrain on j and j' to be equal, while they can happen to coincide though.

Therefore approach A ignores completely the hierarchical structure both in the estimation of the propensity score and in the implementation of the matching. Approach A can succeed in balancing individual-level variables X, but is not meant to build a balanced matched dataset also with respect to Z. In this case, if we do not include all relevant confounders at the cluster-level, Z, in the propensity score and obtain a good balance of all of them, an ATT estimator based on the PSM approach A will be biased.

Approach B deals with this problem by matching units within clusters. In this case, the model used for the estimation of propensity scores is still model (1) and thus it ignores the clustering. However, the matching is implemented within clusters in the sense that the algorithm applies the procedure (2) only to units belonging to the same cluster, i.e., forcing $j = j'$. Therefore:

$$A_{rj} = \left\{ kj \in I_0 : \hat{e}_{kj} = \min_{kj \in I_0} |\hat{e}_{rj} - \hat{e}_{kj}| < 0.20\hat{\sigma}_e \right\}. \tag{5}$$

As in approach A, the matched dataset is created according to (3) and ATT is estimated using (4). Within-cluster matching automatically guarantees that all cluster-level variables (measured and unmeasured) are perfectly balanced. This can come to a cost. Control units to be matched with treated units are only searched within the same cluster. Therefore, it could be that we do not find a control matched unit that we would find in other clusters. So, we may have many more unmatched treated units than with approach A.

To avoid these problems and combine the benefits of approaches A and B, approach C starts by searching control units within clusters (according to (5)). If none is found, control units are searched in other clusters (according to (2)). This approach improves the balancing of cluster-level variables with respect to approach A and avoids the lost of additional units of approach B.

In alternative to exploiting the hierarchical structure in the implementation of the matching, approaches D and E take it into account when modelling the propensity score. In particular, approach D and E use random- or fixed- effects, respectively, to represent unmeasured cluster-level variables. A random-effects logit model can be represented as:

$$logit(e_{ij}) = \alpha_j + X_{ij}\beta, \tag{6}$$

where $\alpha_j \sim N(\alpha_0, \sigma^2_\alpha)$. To estimate propensity scores from (6), predictions of cluster effects α_j are plugged in together with estimates of covariates coefficients.

A fixed-effects logit model differs from (6) for the fact that α_j are not represented by realizations of a random variable but they are fixed intercepts for each cluster that can be estimated by including a set of binary indicators for $J-1$ clusters. After propensity scores have been estimated, both approaches D and E implement the matching on the pooled dataset (procedure (2)).

Arpino and Mealli [8] showed that PSM using random- or fixed-effects models gave similar results in terms of bias reduction and mean squared error of ATT estimators in the presence of unmeasured cluster-level variables and small-size clusters. We consider these approaches also in the case of a strongly unbalanced dataset where some clusters are big and others are small. We compared these approaches with the naive method A and with the pure and preferential within-cluster matching (B and C).

4 Simulation Studies

In this section we describe our simulation experiments aimed at comparing the performance of the different matching strategies described above in the presence of unobserved confounders at the cluster-level.

Set-up

We designed our simulation experiments to mimic the observed data in several respects. First, we kept the same data structure observed in our dataset, i.e. the same number of clusters (hospitals) and the same clusters' sample sizes (see Table 1). In this way, in our simulations we consider a realistic case with a strongly unbalanced structure where some clusters are big and others have small sample sizes. Second, instead of generating values of covariates as realizations of random variables as typically done in simulation studies, we used the same covariates distribution as observed in the dataset. The only exception was for a cluster-level variable, Z, that we introduce to explore the confounding effect at the cluster (i.e., hospital) level. Finally, the coefficients of individual-level covariates in the true models generating the treatment and the outcome were set to values similar to observed coefficients estimated on the real data.

Given the complete set of covariates (X, Z) the probability of being treated was generated according to:

$$e_{ij} = 1/[1 + exp(\beta_0 + \beta_1 X_{1ij} + \cdots + \beta_k X_{kij} + \beta_{k+1} Z_j)] \tag{7}$$

and the outcome was generated by the following model:

$$P(Y_{ij} = 1) = 1/[1 + exp(\gamma_0 + \gamma_1 X_{1ij} + \cdots + \gamma_k X_{kij} + \gamma_{k+1} Z_j + \alpha T_{ij})], \quad (8)$$

where $\boldsymbol{\beta} = [\beta_0, \cdots \beta_{k+1}]$ and $\boldsymbol{\gamma} = [\gamma_0, \cdots \gamma_{k+1}]$ are the vectors of coefficients, $X = (X_1, \cdots X_k)$ is the set of observed individual-level confounders and Z is the cluster-level confounder. Values of Z are generated as realizations of a normal variable with $\mu_Z = 0$ and $\sigma_Z = 0.25$, which is equal to the average standard deviation of the observed confounders. The coefficients $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and α in the equations (7) and (8) were set to values approximately equal to the corresponding coefficients estimated on the real data using logistic regression models without the cluster-level variable Z (see the Appendix for the exact values). The variable Z was introduced in the true data generating models (7) and (8) but it was omitted in the implementation of the PSM strategies described in the previous section to mimic the presence of an unobserved cluster-level confounder. We considered different strengths of the unobserved confounder variable by fixing $\gamma_{k+1} = 0.4$ while β_{k+1} was varied in the set $\{0.2, 0.4, 0.8\}$, corresponding to a small, mild and strong confounding effects (as compared to the magnitude of the coefficients of the others covariates).

To gain further understanding on the performance of the different considered PSM procedures, we expanded the simulation study in two ways:

- varying the coefficients of the individual-level covariates across a subgroup of hospitals,
- fragmenting the data set into a higher number of smaller hospitals.

In the first variation of the basic simulation set-up, the magnitude of the coefficients of the individual-level variables was increased by 25% or 50% in the true model generating the treatment (equation 7) in a randomly chosen subset of 10 hospitals. All other conditions are kept as in the basic simulation set-up. This simulation study is motivated by the fact that in some hospitals some individual-level characteristics may be more important for the decision of using caesarean section than in others. Introducing stronger coefficients in some hospitals implies that the distribution of individual-level variables become more unbalanced in those hospitals between the groups of treated and control units. These varying coefficients by clusters could be modelled trough the inclusion of random slopes (in the random-effect model of procedure D) or interactions between individual-level covariates and hospital indicators (in the fixed-effects model of procedure E). However, in the context of a practical application it may be difficult to know a-priori which covariates should be allowed to have different slopes by clusters and it may be computationally cumbersome to include many random slopes and cross-level interactions [35]. To assess the consequence of ignoring the varying coefficients by clusters additionally to the unobserved hospital-level confounder, we apply the procedures A-E without modifications. Methods B e C should partially deal with varying slopes across clusters "automatically" because matching only or preferentially within clusters should improve the balance of those covariates having stronger effects on T within some clusters.

Finally, to assess the performance of the five considered PSM procedures when all clusters are relatively small, we considered another simulation using a finer partition of the original observations. In practice, we randomly fragmented the original clusters into smaller groups (size between 100 and 200 observations) while keeping constant all other characteristics of the data and of the initial simulation set-up. In the presence of small clusters method B, which insists on matching within clusters only, is expected to perform poorly [10]. In this case, method C is expected to show a much better relative performance. On the other hand, Arpino and Mealli showed that methods D and E performed well also in the presence of small cluster sizes [8].

Under each scenario, 500 datasets were generated from models (7) and (8). For each simulated dataset we employed the PSM methods described in the previous section to obtain a matched subset. The simulation experiments were implemented in R [36]. In particular, for methods A, D and E we obtained the matched subsets using the function `Match` in the package `Matching` [37]. At the time of writing neither this package nor others have an option for implementing within-cluster (B) and preferential within-cluster matching (C) so we programmed a routine that makes use of the `Match` function (the code is available from the authors upon request).

We summarized the results by averaging over the 500 replicates the following metrics calculated on each dataset: the number and the percentage of unmatched treated units, the absolute standardized bias (ASB) of each confounder, the estimated treatment effect ($\widehat{ATT}$), the percent bias of the estimated effect ($\%BIAS$) and the squared error (SE). Note that matching with replacement (with a common caliper and a common propensity score model for all clusters) forces the relation: No. unmatched units (A) $\leq$ No. unmatched units (C) $\leq$ No. unmatched units (B). The ASB was calculated for each confounder as the absolute value of the difference of means between treatment and control group standardized by its standard deviation in the treatment group. The ASB is a measure of covariate balance: a lower ASB indicates that the treatment and control groups are more similar with respect to the given covariate. We report the average ASB for individual-level confounders (X), the ASB for the cluster-level confounder (Z) and for the average ASB of all confounders. The relative bias of each PSM estimators was calculated as

$$|\frac{\widehat{ATT} - ATT)}{ATT}| \cdot 100, \tag{9}$$

where $\widehat{ATT}$ is the average treatment effect on the treated estimated using (4) and ATT is the true value of the average treatment effect on the treated.

Simulation results

Table 2 presents the results of the baseline simulation study introduced in the previous section. We considered three scenarios by varying the effect of the hospital-level unobserved confounder in the true treatment assignment model, $\beta_Z = \{0.2, 0.4, 0.8\}$. For each scenario, we compare the performance of the five PSM strategies described in section 3 (A-E) in terms of unmatched units, balance (ASB), percent bias and mean squared error

(MSE). We also report in the first column the results obtained without any adjustment ("no matching").

First of all, we notice that an unadjusted comparison between treated and control groups' outcomes gives strongly biased estimates (relative bias ranging from 57% to 66%). All PSM methods guarantee a considerable reduction of the bias that tends to increase as the effect of Z increases. However, as expected, only PSM methods that take clustering into account (B, C, D and E) achieve a low bias. Strategy A, that ignores clustering and only adjusts for imbalance in individual-level confounders, shows considerably higher relative biases, especially when the effect of the unobserved hospital-level confounder is the strongest (bias = 24%). The drawback of this method is highlighted by the ASB estimates. Method A always yielded the lowest ASB for individual-level confounders, X, but was, of course, not able to improve the balance of the cluster-level variable, Z.

On the other extreme, method B forces matched pairs to belong to the same cluster. The within-cluster matching guarantees a perfect balance of the unobserved cluster-level variable, Z, at the expense of having only a slightly worse balance of individual-level variables as compared to method A (and also to the other methods). Similarly, method C, that first tries to match units within-cluster and only when this is not possible searches in other clusters, also attains a very good balance of both the cluster-level unobserved confounder and individual-level (observed) confounders. These two methods are those that reduce the bias the most: the relative bias for methods B and C is always lower than 4% while for methods D and E, that use a random- or fixed-effects logistic regression, respectively, to estimate the propensity score, the relative bias is always around 8%.

Method C performs particularly well when the effect of Z is low. Otherwise, the performance of methods B and C is quite similar in terms of relative bias. However, method C has the advantage of reducing the number of unmatched treated units compared to method B. The pure within-cluster matching, in fact, discards on average about 55 units (corresponding to about 1% of the treated units) as compared to less than 1 treated unit that, on average, remains unmatched with method C. Finally, we notice that there is no substantial difference with respect to the variability of ATT estimates as measured by the MSE.

We also considered simulations where the magnitude of the individual-level confounders in the true treatment assignment equation in a subset of 10 hospitals is increased by 25% or 50%. Results appear in the Tables 3 and 4, respectively. The pattern across estimators remain similar to those observed in Table 2. However, higher relative biases are observed for all the PSM approaches. This is reasonable as the propensity score models are now further misspecified. In fact, they assume homogeneous effects of individual-level confounders while the effect of some of them varies across hospitals. Nonetheless, it is reassuring that even in these scenarios, characterized by quite extreme heterogeneous effects for some individual-level confounders, all PSM methods B-E maintain acceptable levels of relative biases. In particular, the pure within-cluster matching (method B) shows a low relative bias (always lower than 5%) in all the scenarios and only a moderate increase in the average number of unmatched treated units as compared to Table 2. This increased number of unmatched units is due to the fact that caliper is set using the propensity

score model estimated on the pooled dataset assuming constant effects of individual-level confounders on the probability of treatment. In those clusters where the effect of individual-level confounders is stronger, the variance of propensity scores is higher than in the pooled sample and is more likely not to find a matched control within the caliper.

Finally, we also assessed the performance of the various PSM estimators when all clusters are relatively small (Table 5). As expected, compared to results in Table 2 the number of unmatched units and the relative bias of the pure within-cluster matching (B) increased conspicuously. Only the performance of the PSM approach based on the random-effect propensity score model (D) is not greatly affected by the different data structure. Method D together with method C are the best performing approaches in these simulations. The preferential within-cluster PSM (Method C) was able to keep the number of unmatched units at very low levels and the relative bias ranged from 6% to 14%. We also notice that the estimator based on a fixed-effects propensity score model (F) is more biased than the one that ignores clustering (A). Method F showed a very high relative bias especially when the effect of Z was strong.

5 Empirical Results

In this section we applied the five PSM approaches to the motivating case study introduced in Section 2. The interest lies in estimating the effect of caesarean section on infants' Apgar score. In our empirical analyses we adjusted for the following covariates: maternal age (five dummy variables: < 20 (reference), $20-24$, $25-29$, $30-35$, > 35 years), maternal education (three dummy variables: less than high school (reference), high school, graduate or higher education), infant birth weight (three dummy variables: < 2500, $2500-4000$ (reference), > 4000 grams), induction of labour (yes=1, no=0), gestational age (in weeks), and presence of a pregnancy-related pathology (binary variable set to 1 if one or more of the following diseases occurred during pregnancy: diabetes mellitus, eclampsia, hypertension, placenta previa). Table 6 reports the standardized mean difference for each covariate calculated on the raw data ("no matching") and then after the implementation of each of the PSM methods. The group of mothers that gave birth with a caesarean section (treated) and those who delivered naturally (control) were quite different with respect to several covariates. For example, the imbalance was relatively high for clinical indications of caesarean section, such as older age, lower gestational age and the presence of pathologies during the pregnancy. Overall, the average ASB was about 14%.

Each of the five PSM approaches we considered reduced considerably the original imbalance. Method A guaranteed the lowest average ASB. However, this method ignores the presence of possible unobserved hospital-level confounders and, as showed by the simulations, gives biased estimate when such confounders exist. Among the other methods, approaches D (random-effect PSM) and E (fixed-effects PSM) gave slightly lower average ASB than the others.

Table 7 shows that the number of unmatched treated units due to the caliper option is very small for the pure within-cluster matching (B) while for the other methods all

treated units can find a matched control unit within the imposed caliper. This is due to the use of the replacement option and to relatively high sample sizes. Moreover, the fact that there were more control units than treated units also favoured the matching.

In Table 8 we show the ATT estimates calculated on the raw data and using the five PSM approaches. All methods consistently report a positive ATT estimate indicating that caesarean section increases the risk of low Apgar scores. The unadjusted difference in the prevalence of low Apgar scores between the treatment and control groups seems to be substantively overestimated. In fact, all five PSM approaches considered gave lower ATT estimates. Interestingly, all methods that adjust for clustering (B-E) gave lower estimates than method A. This may indicate a possible overestimation of the effect of caesarean section when hospital confounding effects are not taken into account.

Even though the estimated ATT is much lower than the one suggested by the unadjusted comparison, it is still important from a substantive point of view: every 1,000 mothers that delivered with caesarean section, the expected number of infants with low Apgar score is increased, compared to the case they had delivered naturally, by about 2 to 3 units depending on the considered PSM estimator.

6 Summary and concluding remarks

Propensity score analyses have been typically considered in the context of unstructured data. In this paper we considered propensity score matching (PSM) for clustered data. Inspired by a real case study regarding estimation of the effect of caesarean section on the Apgar score with data on mothers nested within hospitals, we evaluated via Monte Carlo simulations different strategies to deal with the presence of unobserved confounders at the cluster-level. We assessed two different types of approaches. We considered estimation of the propensity score model using random- or fixed-effects logistic regression. Additionally, we considered matching on a single-level propensity score within-clusters. We also proposed a less extreme variant of the latter approach, that we labelled as preferential within-cluster matching. This method first searches for control units to be matched to treated units within the same cluster. If matching is not possible within-cluster, then the algorithm searches in other clusters.

Differently from Arpino and Mealli [8], our baseline simulations considered a strongly unbalanced structure where some clusters were big enough to justify within-cluster matching, while others had small sample sizes. We confirmed on this type of structure the ability of random- and fixed-effects propensity score models to reduce bias due to omitted cluster-level confounders when used to implement matching. However, we found that the pure (B) and preferential (C) within-cluster matching approaches perform better. In particular, method C was the best method when the strength of the omitted confounder was low or medium while method B had the best performance in the case of a strong omitted confounder. Method B demonstrated to be highly robust to the introduction of varying

coefficients of individual-level confounders across clusters but did not preform well when all clusters were small. In this case, method C worked much better.

Another important aspect to consider when implementing PSM is that a high proportion of unmatched units may imply an implicit change in the estimand [12]. As expected, our simulations showed that imposing a pure within-cluster matching may produce a high proportion of unmatched units, especially when clusters sizes are small. Moreover, we noticed (analyses not reported but available upon request) that the number of unmatched units was not proportional to the cluster size, because the covariates distribution may vary across hospitals. Therefore, method B can imply a change in the estimand not only because some treated units can remain unmatched but also because the data structure can be altered by changing the relative sizes of the different clusters. The preferential within-cluster matching (C) offers a solution to this problem and its performance was good in all scenarios both in terms of relative bias and number of unmatched units.

In summary, our study confirmed the importance of considering the clustering structure in the estimation of causal effects via propensity score matching. Our simulations offer several insights to applied researchers for the choice of the PSM approach in the presence of clustered data. The choice among the methods should mainly depend on the data structure. We showed that when all or at least the majority of clusters' size are big, an effective approach consists in implementing the matching within clusters. When cluster sizes are small, a PSM approach based on a random-effects propensity score model may be a good option. Finally, the preferential cluster matching approach, combining the advantages of within- and between-cluster matching, showed a relatively good performance both in the presence of big and small clusters and it was often the best method.

In our simulations, we assumed that the treatment assignment mechanisms followed a parametric logistic model. We also used a simple data generating process where covariates were included only additively in the linear predictor. In the practice, the true propensity score model may take a complicated form and estimated propensity score model may be misspecified. Machine learning techniques have been proposed as a more flexible way of estimating propensity scores than parametric models [38]. The application of these methods in the context of clustered data structures is an interesting avenue for future research. Finally, it is worth noting that regression adjustment for the propensity score is an alternative to matching, weighting and stratification that has been considered for unstructured data [39]. Its application in the case of multilevel data structures can be considered in future studies.

Acknowledgements

We would like to thank the Autonomous Region of Sardinia for providing the anonymized data used in the empirical application. We are grateful for comments received by participants to the 21^{st} International Conference on Computational Statistics (Geneva) and seminar participants at the Department of Economics and Statistics "Cognetti de Martiis" of the University of Torino (Italy). All errors and inconsistencies are our own.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41-55.

[2] Austin PC. A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in medicine* 2008; **27(12)**: 2037-2049.

[3] Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Statistics in medicine* 2002, **21(21)**: 3291-3315.

[4] Snijders T, Bosker R. *Multilevel Analysis An Introduction to Basic and Advanced Multilevel Modeling. Second Edition.* Sage: London, 2012.

[5] Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; 1231-1236.

[6] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 2004; **86**: 4-30.

[7] D?Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group.*Statistics in medicine* 1998; **17(19)**: 2265?8.

[8] Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis* 2011; **55**: 1770 -1780.

[9] Thoemmes FJ, West SG. The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research* 2011; **46(3)**: 514-543.

[10] Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. *Statistics in Medicine* 2013; **32**(19): 3373-3387.

[11] Gayat E, Thabut G, Christie JD, Mebazaa A, Mary JY, Porcher R. Within-center matching performed better when using propensity score matching to analyze multicenter survival data: empirical and Monte Carlo studies. *Journal of clinical epidemiology* 2013; **66(9)**: 1029-1037.

[12] Crump RK, Hotz VJ, Imbens GW, Mitnik O. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009, **96**:187-195.

[13] Huber M, Lechner M, Wunsch C. The performance of estimators based on the propensity score. *Journal of Econometrics* 2013; **175**: 1-21.

[14] Leyrat C, Caille A, Donner A, Giraudeau B. Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in medicine* 2013, **32(19)**: 3357-3372.

[15] Leyrat C, Caille A, Donner A, Giraudeau B. Propensity score methods for estimating relative risks in cluster randomized trials with low incidence of binary outcomes and selection bias. *Statistics in medicine* 2014 (forthcoming)

[16] Apgar V. A proposal for a new method of evaluation of the newborn infant. *Curr. Res. Anesth. Analg.* 1953; **32**: 260-267.

[17] Finster M, Wood M. The APGAR score has survived the test of time. *Anesthesiology* 2005; **102**: 855-857.

[18] European Perinatal Health Report: The health and care of pregnant women and babies in Europe in 2010 (2013) Available at: www.europeristat.com.

[19] American Academy of Pediatrics, Committee on Fetus and Newborn, American College of Obstetricians and Gynecologists and Committee on Obstetric Practice. The Apgar score. *Pediatrics* 2006; **117**:144-147.

[20] Jain L, Ferre C, Vidyasagar D, Nath S, Sheftel D. Cardiopulmonary resuscitation of apparently stillborn infants: survival and long-term outcome. *J. Pediatr* 1991; **118**: 778-782.

[21] Annibale DJ, Hulsey TC, Wagner CL et al. Comparative neonatal morbidity of abdominal and vaginal deliveries after uncomplicated pregnancies *Arch Pediatr Adolesc Med* (1995); **149**(8) : 862-867.

[22] Cepeda MS, Boston R, Farrar JT, Strom, BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology* 2003; *158(3)*: 280-287.

[23] Bragg F, Cromwell DA, Edozien L et al. Variation in rates of caesarean section among English NHS trusts after accounting for maternal and clinical risk: A cross sectional study. *British Medical Journal* 2010; **6**, DOI:10.1136/bmj.c506.

[24] Van den Berg A, Van Elburg RM, Van Geijn HP, Fetter WP. Neonatal respiratory morbidity following elective caesarean section in term infants: a 5-year retrospective study and a review of the literature.*European Journal of Obstetrics & Gynecology and Reproductive Biology* 2001; **98(1)**: 9-13.

[25] Lavender T, Hofmeyr GJ, Neilson JP, Kingdon C, Gyte GM. Caesarean section for non-medical reasons at term. *Cochrane Database Syst Rev* 2012, **3(3)**.

[26] Werner EF, Han CS, Savitz DA, Goldshore M, Lipkind HS. Health outcomes for vaginal compared with cesarean delivery of appropriately grown preterm neonates. *Obstetrics & Gynecology* 2013; **121(6)**: 1195-1200.

[27] Caceres IA, Arcaya M, Declercq E et al. Hospital Differences in Cesarean Deliveries in Massachusetts (US) 2004-2006: The Case against Case-Mix Artifact. *PLOS ONE* 2013; **8(3)**, DOI:10.1371/journal.pone.0057817.

[28] Hughes RG, Hunt SS, Luft HS. Effects of surgeon volume and hospital volume on quality of care in hospitals. *Med Care* 1987; **25**: 489-503.

[29] Berta P, Seghieri C, Vittadini G. Comparing health outcomes among hospitals: the experience of the Lombardy Region. *Health care management science* 2013; **16(3)**: 245-257.

[30] Rubin DB. Discussion of randomization analysis of experimental data: the fisher randomization test by D. Basu. *Journal of the American Statistical Association* 1980; **75**: 591-593.

[31] Dehejia R, Wahba S. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999; **94**: 1053-1062.

[32] Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J* 2009; **51**: 171-184.

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

[34] Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 2008; **22(1)**: 31-72.

[35] DiPrete TA, Forristal, JD. Multilevel models: methods and substance. *Annual Review of Sociology* 1994; **1**: 331-357.

[36] R Core Team (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* URL http://www.R-project.org/.

[37] Sekhon JS. Multivariate and Propensity Score Matching Software with Automated Balance Optimization. *Journal of Statistical Software* 2011; **42(7)**: 1-52.

[38] Lee BK, Lessler J and Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; **29**: 337-346.

[39] Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. *Statistics in Medicine* 2014; **33**: 4053-4072.

Appendix

In this Appendix we report the exact formulas of equations (7) and (8) in section 4. As explained in the text, the coefficients of individual-level covariates in the true models generating the treatment and the outcome were set to values similar to observed coefficients estimated on the real data.

In particular, the probability of being treated (7) was generated according to:

$$\begin{aligned} e_{ij} &= 1/[1 + exp(\beta_0 + \beta_1 X_{1ij} + \cdots + \beta_k X_{kij} + \beta_{k+1} Z_j)] \\ &= 1/[1 + exp(-1.69 + 0.46 \cdot X_1 + 0.54 \cdot X_2 + 0.74 \cdot X_3 + \\ &\quad + 1.15 \cdot X_4 + 0.08 \cdot X_5 - 0.03 \cdot X_6 + 0.33 \cdot X_7 + \\ &\quad + 0.85 \cdot X_8 + 0.04 \cdot X_9 - 0.2 \cdot X_{10} + \\ &\quad + 0.65 \cdot X_{11} + \beta_{k+1} Z)] \end{aligned}$$

where β_{k+1} was varied in the set $\{0.2, 0.4, 0.8\}$. The outcome (8) was generated according to:

$$\begin{aligned} P(Y_{ij} = 1) &= 1/[1 + exp(\gamma_0 + \gamma_1 X_{1ij} + \cdots + \gamma_k X_{kij} + \gamma_{k+1} Z_j + \alpha T_{ij})] \\ &= 1/[1 + exp(-5.35 + 0.02 \cdot X_1 - 0.12 \cdot X_2 - 0.20 \cdot X_3 + \\ &\quad - 0.47 \cdot X_4 + 0.11 \cdot X_5 + 0.12 \cdot X_6 + 0.58 \cdot X_7 + \\ &\quad - 1 \cdot X_8 - 0.38 \cdot X_9 - 0.2 \cdot X_{10} + \\ &\quad + 0.15 \cdot X_{11} + 0.4 \cdot Z + 0.41 \cdot T)]. \end{aligned}$$

As for the distribution of individual-level covariates we used the empirical distributions observed in the data for the variables we described in section 5. More specifically:

$X_1 \sim$ maternal age (20-24 years)
$X_2 \sim$ maternal age (25-29 years)
$X_3 \sim$ maternal age (30-35 years)
$X_4 \sim$ maternal age ($>$ 35 years)
$X_5 \sim$ maternal education (high school)
$X_6 \sim$ maternal education (graduate or more)
$X_7 \sim$ infant weight ($<$2500 grams)
$X_8 \sim$ infant weight ($>$ 4000 grams)
$X_9 \sim$ induction of labour (yes=1, no=0)
$X_{10} \sim$ gestational age (in weeks)
$X_{11} \sim$ pregnancy related pathologies (yes=1, no=0).

Finally, the cluster-level variable was generated as Z $\sim N(0, 0.25)$.

Table 1: Distribution of births, caesarean sections and low Apgar scores by hospital.

Hospital	No. births	No. caesarean sections	Caesarean sections %	Low Apgar scores ‰
1	2,532	1,166	46.0	16.5
2	1,788	623	34.8	2.7
3	1,687	540	32.0	5.3
4	1,473	632	42.9	14.2
5	1,253	410	32.7	0.7
6	1,197	428	35.7	3.3
7	980	240	24.4	2.0
8	875	238	27.2	5.7
9	529	190	35.9	3.7
10	434	135	31.1	6.9
11	403	164	40.6	0.0
12	396	117	29.5	7.5
13	351	134	38.1	8.5
14	266	74	27.8	7.5
15	208	99	47.5	9.6
16	191	122	63.8	10.4
17	103	40	38.8	9.7
18	50	9	18.0	20
19	32	13	40.6	0.0
20	9	1	11.1	0.0
Total	14,757	5,375	35.0	6.75

Table 2: Simulation results after propensity score matching with replacement.

Metrics	Strategy					
	No matching	A	B	C	D	E
$\beta_Z = 0.2$						
No. unmatched units	0.00	0.62	53.10	0.62	0.82	0.71
% unmatched units	0.00	0.01	0.90	0.01	0.01	0.01
ASB Z	17.90	18.49	0.00	0.25	0.88	1.23
ASB X	13.01	0.95	1.64	1.63	0.93	0.94
ASB All	13.28	1.93	1.55	1.55	0.92	0.96
% Bias	57.42	9.05	3.67	0.61	8.36	8.80
SE	0.0065	0.0035	0.0035	0.0033	0.0034	0.0025
$\beta_Z = 0.4$						
No. unmatched units	0.00	0.81	55.80	0.81	1.45	1.65
% unmatched units	0.00	0.01	0.93	0.01	0.02	0.03
ASB Z	35.72	36.32	0.00	0.35	0.83	0.95
ASB X	12.89	1.04	1.72	1.69	0.99	1.01
ASB All	14.16	3.00	1.62	1.62	0.98	1.00
% Bias	62.76	17.85	2.96	2.62	8.02	8.35
SE	0.0070	0.0038	0.0035	0.0037	0.0036	0.0036
$\beta_Z = 0.6$						
No. unmatched units	0.00	0.93	60.96	0.94	0.84	0.87
% unmatched units	0.00	0.01	0.10	0.01	0.01	0.01
ASB Z	53.03	53.47	0.00	0.62	0.78	0.79
ASB X	12.75	1.15	1.93	1.90	1.08	1.09
ASB All	15.00	4.06	1.83	1.83	1.06	1.07
% Bias	65.88	24.24	2.28	3.78	8.72	7.78
SE	0.0075	0.0042	0.0036	0.0038	0.0037	0.0037

Table 3: Simulation results after propensity score matching with replacement. Coefficients of confounders increased by 25% in a random subset of hospitals.

Metrics	Strategy					
	No Matching	A	B	C	D	E
$\beta_Z = 0.2$						
No. unmatched units	0.00	0.33	54.62	0.33	0.10	0.11
% unmatched units	0.00	0.01	1.37	0.01	0.00	0.00
ASB Z	47.63	48.82	0.00	0.68	0.93	0.97
ASB X	13.36	0.68	1.353	1.35	0.94	0.95
ASB All	15.27	3.94	1.76	1.78	1.14	1.13
% Bias	70.47	22.21	4.60	3.46	10.46	10.72
SE	0.01	0.00	0.00	0.00	0.00	0.00
$\beta_Z = 0.4$						
No. unmatched units	0.00	0.28	59.04	0.28	0.18	0.18
% unmatched units	0.00	0.01	1.42	0.01	0.00	0.00
ASB Z	63.90	64.98	0.00	1.04	0.92	0.88
ASB X	12.78	1.29	1.96	1.94	1.22	1.20
ASB All	15.62	4.83	1.85	1.89	1.21	1.18
% Bias	72.95	29.77	3.50	6.16	9.47	9.31
SE	0.01	0.00	0.00	0.00	0.00	0.00
$\beta_Z = 0.6$						
No. unmatched units	0.00	0.35	63.66	0.35	0.11	0.11
% unmatched units	0.00	0.01	1.45	0.01	0.00	0.00
ASB Z	79.34	80.11	0.00	1.41	0.98	0.96
ASB X	12.18	1.24	2.13	2.10	1.30	1.30
ASB All	15.91	5.62	2.01	2.05	1.28	1.27
% Bias	74.54	34.14	2.74	7.34	9.84	10.29
SE	0.01	0.01	0.00	0.00	0.00	0.00

Table 4: Simulation results after propensity score matching with replacement. Coefficients increased by 50% in a random subset of hospitals.

Metrics	Strategy					
	No matching	A	B	C	D	E
$\beta_Z = 0.2$						
No. unmatched units	0.00	1.37	53.40	1.37	0.00	0.00
% unmatched units	0.00	0.00	1.64	0.00	0.00	0.00
ASB Z	72.34	73.48	0.00	1.22	1.11	1.07
ASB X	12.64	1.58	2.05	2.45	1.37	1.35
ASB All	15.96	5.57	2.36	2.39	1.35	1.34
% Bias	72.71	31.13	3.19	6.46	11.01	10.87
SE	0.01	0.01	0.00	0.00	0.01	0.01
$\beta_Z = 0.4$						
No. unmatched units	0.00	1.32	56.41	1.33	0.00	0.00
% unmatched units	0.00	0.04	1.68	0.04	0.00	0.00
ASB Z	86.36	87.18	0.00	1.56	1.19	1.21
ASB X	11.90	1.40	2.68	2.63	1.41	1.42
ASB All	16.04	6.17	2.53	2.57	1.40	1.40
% Bias	74.35	36.15	3.23	7.67	8.84	9.11
SE	0.01	0.01	0.00	0.00	0.01	0.01
$\beta_Z = 0.6$						
No. unmatched units	0.00	1.11	62.18	1.10	0.00	0.00
% unmatched units	0.00	0.03	1.74	0.03	0.00	0.00
ASB Z	99.43	99.91	0.00	1.98	1.35	1.36
ASB X	11.17	1.26	2.80	2.75	1.52	1.51
ASB All	16.08	6.74	2.65	2.70	1.51	1.50
% Bias	74.91	39.07	1.51	9.84	8.41	8.34
SE	0.01	0.01	0.00	0.00	0.01	0.01

Table 5: Simulation results after propensity score matching with replacement with small clusters.

Metrics	Strategy					
	No Matching	A	B	C	D	E
$\beta_Z = 0.2$						
No. unmatched units	0.00	0.45	446.61	0.45	0.64	0.00
% unmatched units	0.00	0.01	7.76	0.01	0.01	0.00
ASB Z	19.43	20.09	0.00	2.09	0.91	19.04
ASB X	13.36	1.30	1.84	1.81	1.14	1.13
ASB All	13.39	1.94	2.46	2.37	0.85	1.80
% Bias	60.84	12.51	18.97	6.15	10.28	19.38
SE	0.01	0.00	0.00	0.00	0.00	0.00
$\beta_Z = 0.4$						
No. unmatched units	0.00	0.58	464.98	0.58	1.05	0.02
% unmatched units	0.00	0.01	8.01	0.01	0.02	0.00
ASB Z	37.60	39.32	0.00	4.54	0.90	35.76
ASB X	12.74	0.85	2.59	2.37	0.92	0.92
ASB All	14.12	2.99	2.44	2.49	0.92	2.86
% Bias	67.14	22.50	19.00	10.33	10.63	27.85
SE	0.01	0.00	0.00	0.00	0.00	0.00
$\beta_Z = 0.6$						
No. unmatched units	0.00	0.54	497.28	0.54	0.88	0.12
% unmatched units	0.00	0.01	8.48	0.01	0.02	0.00
ASB Z	54.63	57.20	0.00	7.22	0.95	51.42
ASB X	12.32	0.85	2.56	2.33	1.08	0.91
ASB All	14.67	3.98	2.41	2.59	1.07	3.72
% Bias	71.59	31.20	19.39	13.72	11.11	34.78
SE	0.01	0.00	0.00	0.00	0.00	0.00

Table 6: Standardized mean differences of covariates between treated and control groups.

Variable	No matching	A	B	C	D	E
Maternal Age (years)						
< 20	-14.94	-0.53	0.78	0.78	-1.35	-1.01
20-24	-12.64	0.70	0.74	0.60	0.02	-0.93
25-29	-15.04	0.83	2.49	2.53	-0.65	0.33
30-35	-6.11	-0.67	1.08	1.08	0.80	1.98
> 35	26.67	-0.16	-3.43	-3.39	0.01	-1.48
Maternal Education						
Less than High School	-2.53	-1.68	-2.63	-2.91	-3.57	-2.88
High School	0.57	1.14	0.44	0.54	2.52	1.03
Graduate or more	2.80	0.32	2.61	2.77	1.69	0.99
Missing	-0.06	0.46	0.57	0.68	-0.43	1.94
Infant Weight (grams)						
< 2500	21.50	0.42	4.39	4.30	0.74	0.89
2500-4000	-23.90	3.00	-5.52	-5.32	-0.08	-0.10
>4000	9.14	1.34	3.00	2.79	-0.10	-1.20
Labor Induction	-5.04	-3.68	-2.00	-2.07	-3.29	-2.29
Gestational Age	-32.44	2.15	2.46	2.71	2.13	2.70
Pathology during pregnancy	20.76	1.21	5.03	5.55	0.94	2.38
Average Standardised Bias	13.94	1.21	2.48	2.54	1.28	1.48

Table 7: Number of unmatched treated units by hospital.

Hospital	No. births	No. caesarean sections	No. of unmatched treated units				
			A	B	C	D	E
1	2,532	1,166	0	0	0	0	0
2	1,788	623	0	0	0	0	0
3	1,687	540	0	2	0	0	0
4	1,473	632	0	15	0	0	0
5	1,253	410	0	0	0	0	0
6	1,197	428	0	1	0	0	0
7	980	240	0	4	0	0	0
8	875	238	0	0	0	0	0
9	529	190	0	3	0	0	0
10	434	135	0	5	0	0	0
11	403	164	0	3	0	0	0
12	396	117	0	0	0	0	0
13	351	134	0	1	0	0	0
14	266	74	0	0	0	0	0
15	208	99	0	4	0	0	0
16	191	122	0	7	0	0	0
17	103	40	0	0	0	0	0
18	50	9	0	1	0	0	0
19	32	13	0	1	0	0	0
20	9	1	0	0	0	0	0
Total	14,757	5,375	0	47	0	0	0

Table 8: Unadjusted and adjusted (PSM strategies A-E) empirical analyses of the effect of caesarean section on the Apgar score

PSM approach / Metrics	No matching	A	B	C	D	E
No. unmatched units	0.00	0.00	47	0.00	0.00	0.00
ASB	14.80	1.21	2.48	2.53	1.28	1.48
$\widehat{ATT}$	5.75	2.80	2.62	2.25	2.68	2.07