"Clinical Stability" and Propensity Score Matching in Cardiac Surgery: is the clinical evaluation of treatment efficacy algorithm-dependent in small sample size settings?

Daniele Bottigliengo[1], Aslıhan Sentürk Acar[2], Veronica Sciannameo[1], Giulia Lorenzoni[1], Jonida Bejko[3], Tomaso Bottio[4], Emanuele Cozzi[5], Marta Vadori[5], Jean-Paul Soulillou[6], Jean Christian Roussel[6], Thierry Le Torneau[6], Thomas Senage[6], Rafael Mañez[7,8], Cristina Costa[8], Vered Padler-Karavani[9], Sofia Scali[4], Massimiliano Carrozzini[4], Emilia Fiorello[4], Samuel Fusca[4], Gino Gerosa[4], Ileana Baldi[1], Paola Berchialla[10], Dario Gregori[1]

(1) Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac Thoracic Vascular Sciences and Public Health, University of Padova, Padova, Italy
(2) Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey
(3)Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy
(4) Department of Cardiac Thoracic Vascular Sciences and Public Health, University of Padova, Padova, Italy
(5) Azienda Ospedaliera di Padova, Padova, Italy
(6) Centre Hospitalier Universitaire de Nantes, Chirurgie Thoracique et Cardio-Vasculaire, Nantes, France
(7) Infectious Pathology and Transplantation Division, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Hospitalet de Llobregat, Barcelona, Spain
(8) Intensive Care Department, Bellvitge University Hospital, Hospitalet de Llobregat, Barcelona, Spain
(9) Department of Cell Research and Immunology, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, 69978, Israel
(10) Department of Clinical and Biological Sciences, University of Torino, Torino, Italy

CORRESPONDING AUTHOR: *Ileana Baldi, PhD, Unit of Biostatistics, Epidemiology and Public Health, Department of Cardio-Thorax-Vascular Sciences and Public Health, Via Loredan 18, 35121 Padova, Italy; Tel: +390498275384; Fax: +3902700445089; E-mail: ileana.baldi@unipd.it*

DOI: ***10.2427/13001***
Accepted on December 19, 2019

ABSTRACT

Background: Propensity score matching represents one of the most popular techniques to deal with treatment allocation bias in observational studies. However, when the number of enrolled patients is very low, the creation of matched set of subjects may highly depend on the model used to estimate individual propensity scores, undermining the stability of consequential clinical findings. In this study, we investigate the potential issues related to the stability of the matched sets created by different propensity score models and we propose some diagnostic tools to evaluate them.
Methods: Matched groups of patients were created using five different methods: Logistic Regression, Classification and Regression Trees, Bagging, Random Forest and Generalized Boosted Model. Differences between subjects in the matched sets were evaluated by comparing both pre-treatment covariates and propensity score distributions.

We applied our proposal to a cardio-surgical observational study that aims to compare two different procedures of cardiac valve replacement.
Results: Both baseline characteristics and propensity score distributions were systematically different across matched samples of patients created with different models used to estimate propensity score. The most relevant differences were observed for the matched set created by estimating individual propensity scores with Classification and Regression Trees algorithm.
Conclusion: Clinical stability of matched samples created with different statistical methods should always be evaluated to ensure reliability of final estimates. This work opens the door for future investigations that fully assess the implications of this finding.

Key words: Low sample size; Clinical stability; Propensity Score Models; Propensity Score Matching

INTRODUCTION

In surgical research, assessing the effect of a new intervention or procedure is a challenging task. Randomized Controlled Trials (RCTs), which are considered the gold standard for such evaluations, can be difficult to implement in a field where the randomization of patients may be unfeasible or unethical [1]. Despite several studies have investigated some critical aspects that should be addressed to design high quality randomized trials [2,3], conducting RCTs in surgery research is still difficult, impractical and often generate results of difficult generalization [4]. Indeed, evidence-based medicine is very important for clinical decision-making and observational studies can be a valuable tool to assess the potential benefits of a surgical intervention or procedure [5]. In such situations, the lack of randomization poses several issues in the evaluation of potential differences between compared groups of subjects. In fact, physicians often allocate patients to treatment groups given their pre-operative characteristics, such as age, gender and severity of the diseases. Thus, appropriate methodological approaches capable to account for these issues are needed, to ensure that clinical assessment of surgical effects are less confounded by patient's features.

Propensity Score (PS) methods are some of the statistical approaches that can overcome, or at least minimize, consequences of clinically driven, statistically biased allocation. In fact, they are potentially able to recreate the conditions of Randomized Clinical Trials, balancing on average individual baseline characteristics and thus reducing the risk of confounded estimates [6,7]. Propensity Score Matching (PSM) is probably the most popular PS based method of the surgical literature [8–10]. PSM allows for the construction of a sample where individuals from the treatment group are paired with one (or more) individual from the control group with similar PS values, thus with similar baseline characteristics [7].

Several concerns have been however expressed on the acritical usage of such tool [8,11]. Indeed, inclusion of a given patient in the final sample after matching is based on her/his PS, which is constructed to guarantee the capability of the PS function to minimize bias "on average". In surgery, however, two major issues may emerge during PS analysis. The former is related to small number of enrolled patients, which is often encountered in surgery study, that can make PS estimation and, consequently, matched sample creation unstable. The issue of small sample size in the PS analysis has been addressed in the literature [12,13]. In particular, in the study of Pirracchio and colleagues it was found that classical PS approaches, such as PSM and Propensity Score as Inverse Probability of Treatment Weighting (PS-IPTW) led to substantially unbiased estimates of treatment effect. The latter may arise because, from a clinical point of view, the actual patient can be potentially quite different from the virtual "average" patient, as derived from the PS. Thus, monitoring PS values computed for each individual and the individuals that are included in the matched sample can be of major interest in such situations.

Several algorithms have been proposed in literature to produce such PS function. PS is generally estimated using Logistic Regression (LR), that is a parametric approach. Probit regression and discriminant function analysis are other parametric models that can be used for the PS estimation. Parametric models are constrained to a specified form and model misspecifications may produce biased estimates [14]. In recent years, several Machine Learning Techniques (MLT) have been used as an alternative to LR for the estimation of PS [15]. The term machine learning is used for various computing procedures based on logical or binary operations that learn task from several examples [16]. While a priori model with estimated parameters is assumed for modeling in classical statistical approaches, the relationship between an outcome and predictors is constructed by a learning algorithm in MLT [17].

Taking all the above aspects into consideration, the issue of the clinical stability of the produced PS matched

samples may arise. By "clinical stability" we mean the tendency of patients to be included in the final matched sample regardless of the technique used to produce the PS.

This is relevant to guarantee that the clinical evaluation of the research is based on a coherent set of patients and not to single patients, whose inclusion in the analysis may condition the study conclusion.

The aim of this paper is to discuss this concept of "clinical stability" of the matched samples with reference to different techniques used to estimate individual PS and to propose simple diagnostics to evaluate it. We apply our approach to the data of a cardiac surgical observational study funded by the European Union, Translink, which aimed to determine the possible role of the immune response in patients receiving biological cardiac valve as a cause of tissue degeneration in the medium and long term.

MATERIALS AND METHODS

Cardiac valve replacement is the first line therapy for degenerative heart-valve diseases, as it is associated with advantages in terms of life expectancy and reduction of costs to healthcare systems. There are different types of valves available. They can be classified into two major categories: mechanical valves, made of artificial materials, and biological valves, made of natural fabrics, such as pork valves or bovine pericardium. In general, mechanical valves are used in case of rheumatic fever, that afflicts young population, whereas biological valves are used in case of calcific aortic stenosis, that involves an elderly population [http://www.translinkproject.com/about-translink/heart-valve-pathology-the-translink-target/]. The aim is to investigate the process of valve degeneration, according to the type valve, the clinical outcomes and to analyze clinical and biological implications of biological valve implantation.

A total of 112 patients were enrolled in the study: 89 of them underwent a biological valve replacement and 23 a mechanical one. Patients that were already bearers of biological or mechanical valve had valve replacement surgery at the University Hospital of Padova between the years of 2006-2011. A collaboration in the study and postoperative follow-up were requested from the patients recruited for Translink. After Translink's informed consent had been obtained, patients were invited to perform a test battery to investigate psychological well-being and neuropsychological state, to provide an outline of the person at distance of more than five years from valve surgery.

Comparison between groups of patients was carried out considering mechanical-valve recipients as the treatment group and those with biological valve as the untreated/control group. Individual baseline characteristics that were considered as potential confounders were: age, gender, schooling year, smoking status, obesity, hypertension, diabetes type I and II, dyslipidemia, previous stroke, neurological disorder and atrial fibrillation. Neurological disorder and atrial fibrillation were collected immediately after surgery.

Five different techniques were considered for PS estimation: Logistic Regression (LR), Classification and Regression Tree (CART), Bagged Trees (Bagging), Random Forest (RF) and Generalized Boosted Models (GBM).

LR is the most classical method used to compute individual PS values. It belongs to the class of Generalized Linear Model (GLM) and it is widely employed to model binary outcome. CART, Bagging, RF and GBM belong to the family of Tree-based methods. CART [18] is a simple technique that consists of building a decision tree on a given dataset to make predictions. CART divides the space of the covariates in different distinct regions by a recursive binary splitting. Then, predictions are performed in each region. CARTs are very simple to construct and easy to interpret. However, they often suffer from overfitting, which makes them unreliable in many situations. Bagging, RF and GBM were proposed as improvement to the classical CART. They are ensemble of trees that basically consist of building one decision tree on many bootstrap replicates of the original data and average the predictions of each grown tree. Bagging [19] constructs a series of trees on each bootstrap replicate by considering all the available independent variables, while RF [20] builds every tree considering only a subset of the available covariates. GBM [21] makes prediction by estimating a smooth function of several predictors putting together many simple functions. Each simple function considered by GBM is a weak decision tree, characterized by a simple and limited structure.

CART, Bagging, RF and GBM were proposed in several studies as alternatives to LR to estimate PS [22–24]. Their ability to flexible model how subjects were assigned to each treatment group, accounting for non-linearities and interactions, makes their PS estimates more precise and reliable with respect to LR, especially in situations with several confounders and small sample size.

LR, CART, Bagging, RF and GBM methods were used to estimate propensity scores of being treated with mechanical valve. We included only main effects in logistic regression, did not add any polynomial terms and interaction terms. Each model was tuned using 10-fold cross validation and the model with the parameters giving the highest accuracy was selected. At the end of tuning, Bagging had 25 bootstrap replications; RF had 500 trees to grow with 2 variables randomly sampled as candidates at each split and GBM had 50 trees, interaction depth 1, shrinkage parameter 0.1 and 10 observations in the terminal nodes.

Matching was carried out using nearest neighbor algorithm and performing classical 1:1 matching without replacement, where each treated subject was paired with one control subject from the control group with similar PS value.

Agreement between different PS methods was evaluated by comparing the distribution of PS values

estimated by each statistical technique. The evaluation was conducted on the matched sample to understand if matched subjects from the control group were systematically different given the technique used to estimate PS. We carried on the investigation using several indexes: Cohen's kappa, Spearman's rank correlation coefficient, Standardized Median Differences (SMedianD) and Median Absolute Deviation Ratio (MADR). Each of these indexes was computed comparing pair of techniques, e.g. LR vs RF, LR vs. CART, and so on.

Cohen's kappa [25]methodology, and formulation in areas of laboratory performance, instrument or assay validation, method comparisons, statistical process control, goodness of fit, and individual bioequivalence. In all of these areas, one needs measurements that capture a large proportion of data that are within a meaningful boundary from target values. Target values can be considered random (measured with error is an agreement coefficient between two rates for categorical scales. Here, it is used to evaluate the agreement between the patients with biological valves that were included in each matched set given the technique used to estimate PS. Matching result is coded as 1 if the patient with biological valve is matched with a patient with mechanical valve and 0 if not matched. Cohen's kappa ranges between -1 and +1. The higher the value, the higher the concordance and thus the more similar the matched controls.

Spearman's rank correlation coefficients between the estimated propensity scores that belong to matched patients with biological valves (N=23) are calculated. It ranges between -1 and +1. Lower values denote inverse correlation, whereas positive values denote correlation in the same direction. Thus, values close to -1 denote differences between the matched controls of the compared techniques. Values close to 1 denote instead similar matched controls.

SMedianD and MADR are respectively the standardized difference between median PS values and the ratio of the median absolute deviations from PS estimated by compared techniques. Those were preferred over Standardized Mean Difference (SMeanD) and Variance Ratio (VR) because they are more robust to skewed distributions. SMedianD values close to 0 and MADR values close to 1 denote low differences, thus similarity between matched controls.

The evaluation of the differences between the baseline characteristics of the patients were evaluated in terms of SMeanD and VR, following the suggestion of summarizing balance in the matched set using low dimensional summaries, both before and after matching [26,27]authors usually choose the presented estimates from numerous trial runs readers never see. Given the often large variation in estimates across choices of control variables, functional forms, and other modeling assumptions, how can researchers ensure that the few estimates presented are accurate or representative? How do readers know that publications are not merely demonstrations that it is possible to find a specification that fits the author's favorite hypothesis? And how do we evaluate or even define statistical properties like unbiasedness or mean squared error when no unique model or estimator even exists? Matching methods, which offer the promise of causal inference with fewer assumptions, constitute one possible way forward, but crucial results in this fast-growing methodological literature are often grossly misinterpreted. We explain how to avoid these misinterpretations and propose a unified approach that makes it possible for researchers to preprocess data with matching (such as with the easy-to-use software we offer.

All the analyses were implemented using R version 3.3.1 [28]. In particular, CART was fitted using *rpart* R package [29] with default parameters. PS estimation with Bagging, RF and GBM were all implemented using *caret* R package [30]

RESULTS

In Table 1, distributions of baseline characteristics of compared groups in the original sample are reported. Obesity and atrial fibrillation were the only characteristics similar in the biological and mechanical valve groups. The other baseline characteristics systematically differed between groups, with SMeanD often far from 0 (more extreme than 0.1 or -0.1). Regarding VR, schooling year presented a value close to 1. Furthermore, SMD for schooling year was far from 0, thus the variable was considered highly different between groups. Propensity score values estimated with LR, Bagging and RF were close to zero. There were 14, 47 and 1 values close to zero estimated by logistic regression, bagging and RF respectively.

Since classification trees produced poor estimates with only 4 unique values for propensity scores, minimum, 1st quartile and median PS values were equal. PS generally had low values (mean is between 0.1682 and 0.2136) in each model.

To check the agreement between patients included in each matched sample with different estimation methods, Cohen's kappa and 95% confidence intervals are given in Table 2. According to kappa values, CART showed the lowest agreement with the other techniques, with values always close to 0.3. This is also in line with evaluation of PS distributions. Overall, there was greater agreement among ensemble of trees techniques. Indeed, the couples Bagging-RF and RF-GBM showed the highest Cohen's Kappa value, i.e. 0.65 and 0.71, respectively.

According to Table 2, highest positive correlations were between Bagging-RF, CART-RF and LR-GBM. Except for the Bagging-RF, all the other methods showed a low to moderate correlation, with values ranging nearly between 0.3 and 0.6.

TABLE 1. Baseline characteristics of patients with biological and mechanical valves. Categorical variables are reported as percentage (number), whereas continuous variables are reported as 1st quartile/median/3rd quartile (number). Standardized mean difference (SMeanD) and Variance Ratio (VR) are the measures employed to evaluate the balance of baseline characteristics in the matched sample.

Variables	Biological Sample (N=89)	Mechanical Sample (N=23)	Combined (N=112)	SMD	VR
Female	23.6% (21)	8.7% (2)	20.5% (23)	-0.3697	-
Schooling year	5/8/13	6.5/13/13	5/8/13	0.4024	1.145
Age (year)	60/67/71	54/59/62	58/66/70	-1.0878	1.747
Smoker	29% (26)	52% (12)	34% (38)	0.4901	-
Obesity	15% (13)	17% (4)	15% (17)	0.0769	-
Hypertension	45% (40)	61% (14)	48% (54)	0.3185	-
Diabetes mellitus-I	9% (8)	0% (0)	7.1% (8)	-0.3494	-
Diabetes mellitus-II	7.9% (7)	0% (0)	6.2% (7)	-0.3248	-
Dyslipidemia	36% (32)	22% (5)	33% (37)	-0.3018	-
Stroke	5.6% (5)	8.7% (2)	6.2% (7)	0.1262	-
Neurological disorder	2.2% (2)	4.3% (1)	2.7% (3)	0.1291	-
Atrial fibrillation	28% (25)	26% (6)	28% (31)	-0.0444	-

TABLE 2. Cohen's Kappa and Spearman Correlation coefficient of PS on the matched samples (95% Confidence Interval)

COHEN'S KAPPA

	M1 (LR)	M2 (CART)	M3 (Bagging)	M4 (RF)	M5 (GBM)
M1 (LR)		0.24 (0.016,0.46)	0.36 (0.14,0.57)	0.53 (0.33, 0.73)	0.53 (0.33, 0.73)
M2 (CART)			0.36 (0.14, 0.57)	0.3 (0.076, 0.52)	0.36 (0.14, 0.57)
M3 (Bagging)				0.65 (0.47, 0.83)	0.53 (0.33, 0.73)
M4 (RF)					0.71 (0.54, 0.88)
M5 (GBM)					

SPEARMAN CORRELATION COEFFICIENT

	M1 (LR)	M2 (CART)	M3 (Bagging)	M4 (RF)	M5 (GBM)
M1 (LR)		0.348 (-0.075, 0.665)	0.365 (-0.055, 0.676)	0.367 (-0.054, 0.677)	0.609 (0.262, 0.816)
M2 (CART)			0.593 (0.240, 0.808)	0.602 (0.253, 0.813)	0.511 (0.126, 0.763)
M3 (Bagging)				0.987 (0.969, 0.994)	0.541 (0.166, 0.780)
M4 (RF)					0.511 (0.179, 0.785)
M5 (GBM)					

In Table 3, SMedianD and MADR values of biological group PS distributions in the matched sets are reported. Regarding SMedianD, the pairs LR-GBM and Bagging-RF showed the values closest to 0, while the couple RF-GBM showed the highest difference (SMedianD = 0.6468). All MADR values were often far from 1, except for the comparison RF-Bagging, with MADR value close to 1.33. The highest difference was observed for LR-RF (3.621) and the couples LR-CART, CART-Bagging, CART-RF and CART-GBM, which showed a MADR almost equal to 0, suggesting some extreme discrepancies between the MADs of the considered techniques.

In Table 4, patient's characteristics are reported for each patient. The column labelled as "Number of inclusions" identifies the number of times the patient was included in the final matched sample created using different techniques to estimate PS. Moreover, variables indicating if the patient was included in the matched set are provided. These data showed that the sets of patients were highly different according to the estimation methods.

TABLE 3. SMedianD and MADR between PS values of subjects from the biological group in the matched sets. Comparison is carried out on couple of techniques used to estimate PS.

SMEDIAND

	M1 (LR)	M2 (CART)	M3 (Bagging)	M4 (RF)	M5 (GBM)
M1 (LR)		-0.2165	-0.3056	0.3476	0.1757
M2 (CART)			-0.2323	0.3353	-0.5536
M3 (Bagging)				0.0439	-0.5812
M4 (RF)					0.6468
M5 (GBM)					

MADR

	M1 (LR)	M2 (CART)	M3 (Bagging)	M4 (RF)	M5 (GBM)
M1 (LR)		0.000	0.368	3.621	0.635
M2 (CART)			0.000	0.000	0.000
M3 (Bagging)				1.333	0.580
M4 (RF)					2.299
M5 (GBM)					

In Table 5, the distribution of individual characteristics in the matched samples are reported. Balance was evaluated both numerically, using SMeanD and VR, and graphically, as reported in Figure 1. All the matched sets created with the considered techniques presented residual imbalance. Regarding LR, a good balance was reached for Schooling year, Obesity, Hypertension, Stroke and Atrial Fibrillation, which showed SMeanD values close to 0 and VR value close to 1, while the others remained unbalanced. Situations were more complex with CART and Bagging methods. In the first case, a good balance was achieved only for Schooling year and Stroke while in the second case only Schooling year and Obesity were balanced after matching, as suggested by lower SMeanD values and VR close to 1. Using GBM to estimate PS, a good balance was reached for Schooling year, Dysplipidemia, Stroke, Neurological disorder and Atrial fibrillation, whereas the use of RF led to a good balance for Smoker, Obesity, Neurological disorder and Atrial fibrillation, with SMeanD values close to 0 and VR values close to 1. Overall, the matched samples produced with PS estimated by LR, GBM and RF were the more balanced and Schooling year, Stroke, Obesity and Atrial fibrillation were the baseline characteristics more similar between matched treated and controls in the final sets of patients.

DISCUSSION

Here, we considered the problem of "clinical stability" that could arise when PSM is performed in situations where sample size is small and the number of subjects in the compared groups is highly different. We evaluated such concept on a study from cardiac surgical research, where subjects with mechanical and biological valves were compared. We compared five different statistical techniques to estimate PS values. Comparison was made on distributions of PS values estimated by each technique on the final matched samples. We used several diagnostic tools to evaluate if the presence of an individual from the biological group was technique-dependent or not.

By comparing PS distributions estimated by different techniques we found some complexity. Three of the five techniques estimated PS values equal to 0 for many subjects. Thus, for these techniques, one of the conditions of ignorable treatment assignment, i.e. each subject has a nonzero probability of being assigned to the treatment group [31], failed and the following estimation of treatment effect conditioned on PS may not be unbiased. Moreover, CART estimated only 4 different values of PS. This may be then reflected in suboptimal matching, where residual imbalance in baseline characteristics may still be present after matching. Overall, all the distributions of PS seemed to be quite different between techniques, suggesting that every subject may have a technique-dependent PS value.

Analyzing the different matched samples, we found many flags that question the concept of clinical stability. First, except for RF and GBM, the matched samples created with different techniques showed very low agreement, suggesting that biological groups in each matched set may be different from each other. Regarding Spearman's correlation coefficients, an high correlation was only observed for the couple Bagging-RF, whereas all the other couples did not show any consistent correlation. Moreover, SMedianD and MADR values showed persistent differences between PS distributions between biological groups of each matched sample. Among all the considered measures, Cohen's Kappa may be the one

TABLE 4. Patient level details with matching results based on each estimation method and covariate information. The column labelled as "Number of inclusions" identifies the number of times the patient was included in the final matched sample created using different techniques to estimate PS.

ID	Number of inclusions	gender	schooling	age	smoker	obesity	hypert.	diabetI	diabetII	Dyslip.	stroke	neu. dis.	atr.fib.
1	0	1	5	79	0	0	1	0	0	0	0	0	1
2	2	1	13	66	1	1	0	0	0	0	0	0	0
3	0	2	5	71	1	0	0	1	0	1	0	0	0
4	1	2	13	75	1	0	1	0	0	0	0	0	0
5	3	1	8	56	0	0	1	0	0	0	0	0	0
6	0	2	5	69	0	0	1	0	0	0	0	0	1
7	2	1	11	58	0	0	1	0	0	1	0	0	1
8	1	1	5	66	0	0	1	0	0	0	0	0	1
9	0	1	8	69	0	0	0	0	0	0	0	0	0
10	5	1	13	55	0	1	1	0	0	0	0	0	1
11	3	1	13	50	0	0	0	0	0	1	0	0	0
12	1	2	8	46	0	1	0	0	0	0	0	0	0
13	2	1	13	66	0	0	0	0	0	0	0	1	0
14	1	1	5	65	0	0	1	0	0	0	0	0	0
15	4	1	10	47	1	1	1	1	0	1	0	0	0
16	2	1	5	67	1	0	0	0	0	0	0	0	0
17	4	1	13	58	0	0	0	0	1	0	0	0	0
18	0	1	8	69	0	0	1	0	1	1	0	0	0
19	0	1	13	66	0	0	1	0	0	0	0	0	1
20	0	1	13	72	1	1	1	0	0	1	0	0	0
21	0	1	5	74	1	0	0	0	0	1	0	0	0
22	0	1	5	60	0	0	0	0	0	0	0	0	0
23	3	1	8	65	1	1	1	0	0	0	0	0	0
24	3	2	5	62	1	1	0	0	0	1	0	0	0
25	1	1	11	61	0	0	0	0	0	0	0	0	0
26	3	2	8	55	1	0	0	0	0	0	0	0	0
27	1	1	5	76	1	0	1	0	0	1	1	0	0
28	0	2	8	70	0	0	0	0	0	0	0	0	0
29	0	1	8	71	1	0	0	0	0	1	0	0	1
30	0	2	5	74	0	0	0	0	0	0	0	0	1
31	0	2	5	73	0	0	0	0	0	0	0	0	1
32	1	1	11	55	0	0	0	0	0	1	0	0	0
33	4	1	13	56	0	0	1	0	0	1	1	0	0
34	0	1	5	80	0	0	1	0	0	0	0	0	1
35	0	1	5	71	0	0	0	0	0	0	0	0	1
36	4	2	8	42	0	1	1	0	0	0	0	0	0
37	4	1	10	53	1	0	0	0	0	0	0	0	0
38	0	1	8	73	1	1	0	0	0	1	0	0	0
39	1	1	7	58	0	0	1	1	0	1	0	0	1
40	4	1	18	63	1	0	1	1	0	1	0	0	0

of more interest. Indeed, it can provide more information from a clinical point of view by measuring the similarity of the matched sets created with different techniques and, consequently, informing the physicians on how many times each patient was included in the final sample. Spearman's coefficient, SMedianD and MADR can be valuable the PSM analysis step to inspect the eventual discrepancies between the PS estimated by each technique.

Evaluating the final matched samples, we found that residual imbalance in baseline characteristics was still present in all reconstructed sets created with different PS. More balanced sets were observed when LR, GBM and RF were used to estimate PS and only a subset of covariates often reached balance in all the final matched samples. Indeed, final outcome analysis should be conducted with proper methods that takes into account the role of

TABLE 4 (CONTINUED). Patient level details with matching results based on each estimation method and covariate information. The column labelled as "Number of inclusions" identifies the number of times the patient was included in the final matched sample created using different techniques to estimate PS.

ID	Number of inclusions	gender	schooling	age	smoker	obesity	hypert.	diabetI	diabetII	Dyslip.	stroke	neu. dis.	atr.fib.
50	1	1	8	68	1	0	0	0	0	0	0	0	1
51	2	1	18	77	1	0	1	0	0	0	0	0	1
52	5	1	18	54	1	0	0	0	0	1	0	0	0
53	0	1	5	69	0	0	0	0	0	0	0	0	0
54	0	1	8	71	0	0	1	0	0	0	0	0	0
55	1	1	5	61	0	0	0	0	1	0	0	0	0
56	4	1	10	35	1	0	0	0	0	1	0	0	0
57	0	1	13	71	0	0	0	0	0	0	0	0	0
58	0	1	5	70	0	0	0	0	0	0	0	0	1
59	3	1	8	56	0	0	1	0	0	0	0	0	1
60	5	2	5	60	1	0	0	0	0	0	0	0	0
61	0	2	5	67	0	1	0	0	0	1	0	0	0
62	0	1	5	75	0	0	0	0	0	1	0	0	0
63	1	2	16	64	0	0	0	0	0	0	0	0	0
64	2	2	16	62	1	0	0	0	0	0	0	0	0
65	0	1	14	62	0	0	0	0	0	1	0	0	0
66	0	1	5	72	0	0	1	1	0	1	0	0	1
67	0	2	5	74	0	1	1	0	0	1	0	0	1
68	4	1	18	61	0	0	0	0	0	0	0	1	1
69	2	1	8	53	0	0	0	0	0	0	0	0	0
70	0	1	8	71	0	0	0	0	0	0	1	0	0
71	0	2	5	66	0	0	1	0	1	0	0	0	0
72	0	1	5	76	0	1	0	0	0	1	0	0	0
73	0	1	5	68	0	0	1	0	0	1	0	0	0
74	0	2	5	73	0	1	1	0	0	1	0	0	1
75	1	1	17	66	1	0	1	0	0	1	0	0	1
76	0	2	10	71	0	0	1	1	0	0	0	0	0
77	0	2	17	67	0	0	0	0	0	0	0	0	0
78	1	1	5	63	1	0	0	1	1	0	1	0	0
79	2	1	8	59	0	0	0	0	0	0	0	0	0
80	0	1	5	67	0	0	1	0	0	1	0	0	0
81	0	1	11	68	0	0	0	0	0	0	0	0	1
82	5	1	13	62	0	0	1	0	0	0	0	0	0
83	1	1	13	72	1	0	1	0	0	1	0	0	0
84	0	1	3	74	0	0	1	0	0	1	0	0	1
85	0	1	5	70	0	0	1	0	1	0	0	0	0
86	0	2	5	70	0	0	1	0	0	0	0	0	0
87	5	1	13	57	1	0	1	0	0	0	1	0	0
88	0	1	5	70	0	0	0	0	0	0	0	0	0
89	0	1	8	68	0	0	0	0	0	0	0	0	0

potential confounders that baseline covariates may still play [32,33].

All these findings seem to point the importance of the concept of clinical stability in situations where the analyzed sample has few subjects and the number of subjects between compared groups is very different. The high number of estimated PS values exactly equal to 0 raise some issue on the consequent estimation of treatment effect. The complex behavior of estimated PS suggests that many techniques suffered from overfitting, an issue that is difficult to face, even for ensemble of trees methods, and could raise some inaccuracies in consequent estimations when few subjects are enrolled in the study and many potential confounders are present. In such situations, a

TABLE 5. Distribution and balance assessment of baseline characteristics in the matched samples. Categorical variables are reported as percentage (number), whereas continuous variables are reported as 1st quartile/median/3rd quartile (number). SMeanD and VR are the measures employed to evaluate the balance of baseline characteristics in the matched sample.

Mechanical samples (N = 23)	Biological sample (N = 23)																	
	M1 (LR)			M2 (CART)			M3 (Bagging)			M4 (GBM)			M5 (RF)		
Variables		SMeanD	VR		SMeanD	VR		SMeanD	VR		SMeanD	VR		SMeanD	VR
Female	8.7% (2)	17.4% (4)	-0.255	-	21.7% (5)	-0.137	-	13.0% (3)	-0.255	-	21.7% (5)	-0.255	-	13.0% (3)	-0.255	-
Schooling year	6.5/13.0/13.0	8.0/10.0/13.0	0.010	1.15	8.0/11.0/13.0	0.000	1.19	10.0/13.0/13.0	-0.031	1.13	8.0/10.0/13.0	-0.103	1.19	8.0/13.0/13.0	-0.191	1.00
Age (year)	54/59/62	55/58/64	-0.279	2.01	54/58/62	-0.256	1.78	54/60/63	-0.261	2.02	54/56/60	-0.201	1.78	54/57/62	-0.257	1.41
Smoker	52% (12)	43% (10)	0.170	-	39% (9)	0.349	-	57% (13)	0.171	-	43% (10)	0.349	-	57% (13)	-0.086	-
Obesity	17% (4)	17% (4)	0.000	-	22% (5)	0.119	-	17% (4)	0.000	-	17% (4)	0.000	-	22% (5)	-0.107	-
Hypertension	61% (14)	61% (14)	0.000	-	43% (10)	0.436	-	43% (10)	0.529	-	48% (11)	0.258	-	52% (12)	0.172	-
Diabetes mellitus I	0% (0)	0% (0)	-	-	13.0% (3)	-0.536	-	17.4% (4)	-0.427	-	17.4% (4)	-0.536	-	8.7% (2)	-0.635	-
Diabetes mellitus II	0% (0)	0% (0)	-	-	8.7% (2)	-0.635	-	8.7% (2)	-0.427	-	4.3% (1)	-0.427	-	4.3% (1)	-0.427	-
Dyslipidemia	22% (5)	35% (8)	-0.286	-	43% (10)	-0.557	-	30% (7)	-0.286	-	35% (8)	0.000	-	30% (7)	-0.287	-
Stroke	8.7% (2)	8.7% (2)	0.000	-	8.7% (2)	0.000	-	13.0% (3)	-0.137	-	8.7% (2)	0.000	-	8.7% (2)	-0.137	-
Neurological disorder	4.3% (1)	8.7% (2)	-0.172	-	0.0% (0)	0.295	-	8.7% (2)	-0.173	-	4.3% (1)	0.000	-	4.3% (1)	0.000	-
Atrial fibrillation	26% (6)	26.0% (6)	0.000	-	13% (3)	0.326	-	17% (4)	0.207	-	22% (5)	0.000	-	22% (5)	0.000	-

more parsimonious PS model with covariates only related to the outcome may be helpful to produce more stable estimates of PS using different techniques [12].

Our study has some limitations. We evaluated the concept of clinical stability only in one scenario, i.e. our motivating example. In the future, a simulation study will be considered to address the issue under different scenarios, e.g. different sample size, different ratios between number of subjects in the compared groups and different proportions of continuous and categorical variables.

In summary, the concept of clinical stability should be considered when estimating treatment effect using PSM in a setting with small sample size and compared groups with different number of subjects. We suggest the usage of both qualitative and quantitative diagnostic tools to evaluate this phenomenon. Further explorations are needed to implement a robust approach able to face anticipated complexities.

Acknowledgements

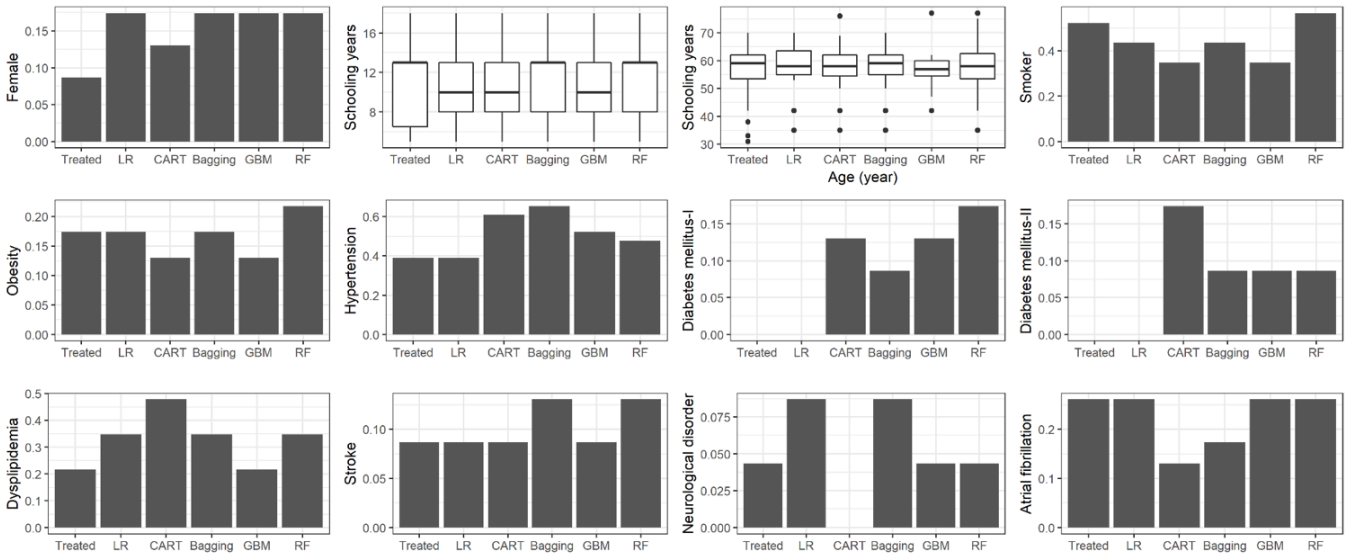
Aslıhan Sentürk Acar has been supported by Scientific Research Projects Coordination Unit of Hacettepe University (project ID: FBI-2017-13640) during her studies at the University of Padova.

This work was supported by the European Union Seventh Framework Program (FP7/2007/2013) under the Grant agreement 603049 for Translink consortium (http://www.translinkproject.com/).

References

1. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ 2002;324(7351):1448–51.
2. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. Trial. 2009;10(1):9.
3. Wu R, Glen P, Ramsay T, Martel G. Reporting quality of statistical methods in surgical observational studies: protocol for systematic review. Syst Rev 2014;3(1):70.
4. Taggart DP. Coronary Artery Bypass Grafting is Still the Best Treatment for Multivessel and Left Main Disease, But Patients Need to Know. Ann Thorac Surg 2006;82(6):1966–75.
5. Ergina PL, Cook JA, Blazeby JM, Boutron I, Clavien P-A, Reeves BC, et al. Challenges in evaluating surgical innovation. The Lancet 2009;374(9695):1097–104.
6. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivar Behav Res 2011;46(3):399–424.
7. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17(19):2265–81.
8. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions

FIGURE 1. Graphical representation of the distributions of baseline characteristics in the matched samples. Bars plots are used to represent percentages of categorical variables, whereas boxplots are used for continuous variables. Discrepancies between baseline covariates in the final sample is carried out by comparing values of the matched treated (labelled as "Treated") with values of the matched controls selected using different models to estimate PS.



for improvement. J Thorac Cardiovasc Surg 2007;134(5):1128-1135.e3.

9. Winger DG, Nason KS. Propensity-score analysis in thoracic surgery: When, why, and an introduction to how. J Thorac Cardiovasc Surg 2016;151(6):1484–7.
10. Hwang ES, Wang X. Value of Propensity Score Matching to Study Surgical Outcomes. Ann Surg 2017;265(3):457.
11. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27(12):2037–49.
12. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163(12):1149–56.
13. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the Propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol 2012 ;12:70.
14. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R [Internet]. New York: Springer-Verlag; 2013 [cited 2018 May 7]. (Springer Texts in Statistics). Available from: //www.springer.com/us/book/9781461471370
15. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol 2010;63(8):826–33.
16. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol 2013;66(4):398–407.
17. Tufféry S. Data Mining and Statistics for Decision Making [Internet]. Wiley.com. 2011 [cited 2018 May 7]. Available from: https://www.wiley.com/en-us/ata+Mining+and+Statistics+for+Decision+Making-p-9780470688298
18. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Taylor & Francis; 1984. 372 p.
19. Breiman L. Bagging Predictors. Mach Learn. 1996;24(2):123–40.
20. Breiman L. Random Forests. Mach Learn 2001;45(1):5–32.
21. Ridgeway G. The State of Boosting. 1999.
22. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004;9(4):403–25.
23. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med 2010;29(3):337–46.
24. Watkins S, Jonsson Funk M, Brookhart MA, Rosenberg SA, O'Shea TM, Daniels J. An Empirical Comparison of Tree-Based Methods for Propensity Score Estimation. Health Serv Res 2013;48(5):1798–817.
25. Lin L, Hedayat AS, Sinha B, Yang M. Statistical Methods in Assessing Agreement. J Am Stat Assoc 2002;97(457):257–70.
26. Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Polit Anal 2007;15(3):199–236.
27. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28(25):3083–107.
28. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/
29. Therneau T, Atkinson B, port BR (producer of the initial R, maintainer 1999-2017). rpart: Recursive Partitioning and Regression Trees [Internet]. 2018 [cited 2018 May 7]. Available from: https://CRAN.R-project.org/package=rpart
30. Wing MKC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: Classification and Regression Training [Internet]. 2018 [cited 2018 May 7]. Available from: https://CRAN.R-project.org/package=caret
31. Rosenbaum PR, Rubin DB. The central role of the propensity score in

observational studies for causal effects. Biometrika 1983;70(1):41–55.
32. Rubin DB, Thomas N. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. J Am Stat Assoc 2000;95(450):573–85.
33. Austin PC. Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching. Stat Methods Med Res 2017;26(1):201–22.