SORT 38 (2) July-December 2014, 271-284

Balancing properties: A need for the application of propensity score methods in estimation of treatment effects

Arantza Urkaregi[1,2], Lorea Martinez-Indart[2,3] and José Ignacio Pijoán[2,3,4]

Abstract

There has been recently a striking increase in the use of propensity score methods in health sciences research as a tool to adjust for selection bias in making causal inferences from observational controlled studies. However, reviews of published studies that use these techniques suggest that investigators often do not pay proper attention to thorough verification of appropriate fulfilment of propensity score adjusting properties. By using a case study in which balance is not achieved, we illustrate the need to systematically asses the accomplishment of the balancing property of the propensity score as a critical requirement for obtaining unbiased treatment effects estimates.

MSC: 62P10, 62J12, 62G10.

Keywords: Propensity score, balancing score, treatment effect.

1. Introduction

In assessing the impact of a clinical intervention an experimental approach through the use of a randomized trial is always regarded as a reference of optimum design leading to the highest quality evidence if properly conducted (D'Agostino and D'Agostino, 2007; Friedman, Furberg and DeMets, 1998). Randomized assignment of alternative interventions minimizes the risk of selection bias (confounding by indication) (Walker, 1996)

[1] Department of Applied Mathematics, Statistics and Operational Research. University of the Basque Country (UPV/EHU). E-mail: arantza.urkaregi@ehu.es

[2] BioCruces Health Research Institute

[3] Clinical Epidemiology Unit-Cruces University Hospital

[4] Network Biomedical Research Centre for Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain

Received: September 2013
Accepted: July 2014

and therefore maximizes internal validity of the causal inferences. Unfortunately, many times ethical, economic or practical reasons impede the use of this experimental design.

The effect of many interventions is instead assessed using observational studies. In non-experimental designs it is necessary to take into account the potential existence of selection bias due to the fact that groups to be compared are not genuinely "comparable" (Grimes and Schulz, 2002). The treatment or intervention any individual receives can be influenced by a mix of measurable and immeasurable factors. We might consider among them, physician preference or belief about the specific effect of a given intervention according to the patient profile, local clinical practice patterns or patient preferences and values. Propensity scores (PS) techniques (Rosenbaum and Rubin, 1983) conform a set of statistical methods devised to minimize this bias.

Although the original paper written by Rosenbaum and Rubin addressed a clinical problem as an example of application, this method has been scarcely used in the health sciences until the last decade. A substantial increase in researchers' interest in and use of this method has been recently detected (Sturmer et al., 2006).

A search of the term "propensity score" in MEDLINE and EMBASE bibliographic databases permit us confirm this tendency and shows that the number of papers that use this approach keeps exponentially increasing (Figure 1).

By using the PS (the conditional probability of receiving the intervention of interest given the pre-treatment individual covariates) we reduce the multidimensionality of the pre-intervention covariate vector to a single number (scalar) that encapsulates all the original information. All individuals with a given PS are expected to have a homogeneous distribution of relevant baseline characteristics, irrespective of whether or not they have received the intervention of interest. Therefore, it is stated that, conditional on these measured pre-intervention covariates, allocation of interventions can be thought

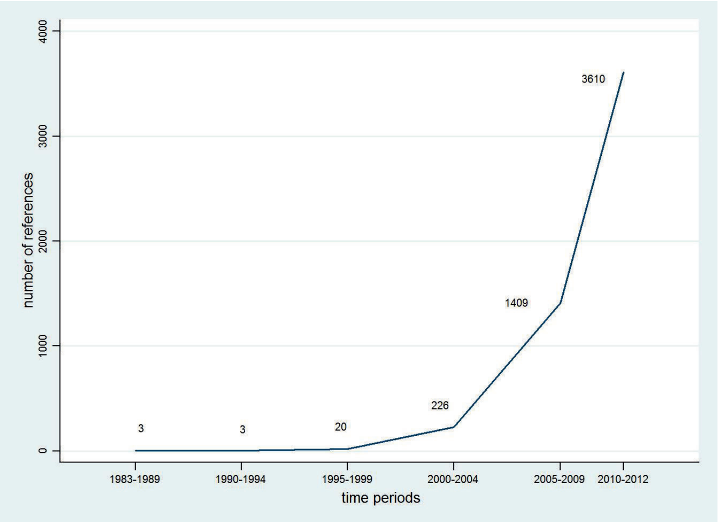

Figure 1: *Time distribution of publications that include the "propensity score" term.*

of as a random process, similar to what happens in a clinical or community trial (Austin, 2011).

One of the theoretical foundations of the adjusting ability of the PS is that it is a balancing score, (Rosenbaum and Rubin, 1983). It implies both, that PS achieves homogeneous covariate distributions between groups and that given the PS, treatment received and covariates are conditionally independent. However, in many cases, the fulfilment of the balancing property of the empirically estimated PS is not systematically assessed (Weitzen et al., 2004). When this is the case there is no real guarantee that the covariates making for the PS are actually adequately balanced; this can, in turn, lead to unfair comparisons. It has been proved (Rosenbaum and Rubin, 1983) that the PS is a balancing score but authors warned that in certain practical conditions that balance cannot be reached. In this article we present a case study in which proper balance could not be achieved because of the highly deterministic nature (lack of significant uncertainty) of the treatment assignment process. If that happens, PS-based methods should not be used to estimate treatment effects.

Methodology

Data description

We analysed data from a clinical cohort of 4,339 neonates with respiratory problems due to prematurity. Some of them were given pulmonary surfactant, a tensioactive substance which improves the mechanics of breathing. Our aim was to estimate the effect of this medical intervention on probability of death during the first 28 postnatal days.

Estimation of PS

PS is defined as the conditional probability of receiving a given treatment conditional on the observed pre-treatment characteristics of the individual. In our step by step PS estimation process every recorded pre-treatment covariate deemed to be important by clinicians with regard to clinical management and treatment of breathing problems in prematurity and/or early prognosis was first pre-selected. Separate bivariate logistic regression models were then fitted to assess the relationship of each pre-selected covariate firstly with treatment received (pulmonary surfactant) and then with outcome (death in the first 28 days after birth). The PS is ultimately estimated from a multivariable binary logistic regression with treatment received as the dependent variable and predictor variables selected from the previous steps. We first included all physician recommended covariates that were shown to be associated with the outcome in previous bivariate models, irrespective of their relationship with the treatment choice (Brookhart et al., 2006). Then, through a manual, step by step, backward approach, variables that were not statistically significant in the multivariable model were removed until the final estimation model was obtained. Predicted probabilities from this model represent the estimated PS.

After PS estimation, we proceeded to check for the existence and pattern of overlapping (common support) in PS values between individuals in comparing groups. PS methods rely on the so-called "counterfactual or potential outcomes" framework (Oakes and Johnson, 2006). For each individual and intervention, there are two potentially observable outcomes: one if she receives the study intervention and another if she does not receive it. The natural effect of the intervention on the subject would be obtained as the difference between these two potential outcomes. In practice, however, only one outcome can be actually observed as the individual either does or does not receive the intervention of interest. Lacking the natural reference for comparison (the same individual in the counterpart situation), we must ensure a proper comparison group exists as a proxy for this unobservable counterfactual experience.

Ensuring that for each selected interval of PS values there are both treated and untreated individuals (Caliendo and Kopeining, 2008) is a required criterion that comparable experience is available, enabling causal inference and estimation of intervention effects. With real data, it is commonplace to find, especially at the tails of the PS distribution, regions where only treated or untreated individuals are found. This finding affects comparability of groups (also referred to as positivity) and produces biased estimates, based partly on extrapolations (Shadish and Steiner, 2010). To appraise the observed degree of overlapping achieved we used descriptive statistics and graphical tools that help inspect the empirical distributions of estimated PS among each treatment group (histograms and box-plots) and took special care in inspecting the tails of the distributions. Additionally nonparametric density estimators (kernel functions) were used to explore and detect potential non-overlapping regions within the whole range of observed PS.

To ensure estimates based on comparable subjects (Pattanayak, Rubin and Zell, 2011), we excluded neonates from the tail areas where there was no overlapping (all untreated neonates whose PS was smaller than the smallest PS in the treated units and all the treated neonates whose PS was larger than the largest PS in the untreated).

Finally balancing properties of the estimated PS were assessed. This is a two-step process: it should be first checked whether the PS is similarly distributed between treated and untreated groups over defined regions across the PS observed range. If this requirement is fulfilled, then assessment of homogeneity of distributions should additionally be performed for each covariate included in the final PS estimation model over the pre-specified regions of observed PS (Adelson, 2013). Only after these two conditions are met we can accept the empirically estimated PS is working adequately as a balancing score.

To do it in practice we started splitting the observed range of PS values into five blocks of equal size (quintiles) (Rosenbaum and Rubin, 1984) and statistically tested the balance of PS between treated and untreated within each block by the use of nonparametric tests (Kolmogorov-Smirnov test). A statistical significance threshold of 0.01 was chosen to account for the chance effect of multiple comparisons (Benjamini and Hochberg, 1995). If balance in PS was not achieved in a specific block, it was further subdivided into two new blocks of the same size and PS balance between the groups was

Table 1: *Pre-selected variables and association with treatment and outcome. DR: delivery room. NEC: necrotizing enterocolitis. RDS: respiratory distress syndrome. PIVH: peri/intraventricular haemorrage. PDA: patent ductus arteriosus. Only statistically significant or marginally significant associations are shown.*

Variables associated with treatment	**Variables associated with outcome**
Gestational age (week and day) ($p < 0.001$)	Gestational age (week and day) ($p < 0.001$)
Mode of delivery ($p < 0.001$)	Mode of delivery ($p < 0.001$)
Gender ($p < 0.001$)	Gender ($p = 0.033$)
Multiple Birth ($p < 0.001$)	—
Apgar test score at 5 minutes ($p < 0.001$)	Apgar test score at 5 minutes ($p < 0.001$)
Endotracheal intubation in DR ($p < 0.001$)	Endotracheal intubation in DR ($p < 0.001$)
Adrenaline /Epinephrine in DR ($p < 0.001$)	Adrenaline /Epinephrine in DR ($p < 0.001$)
Cardiac Compression in DR ($p < 0.001$)	—
Prenatal corticosteroid use ($p < 0.001$)	Prenatal corticosteroid use ($p < 0.001$)
Conventional Ventilation after leaving DR ($p < 0.001$)	Conventional Ventilation after leaving DR ($p < 0.001$)
High Frequency Ventilation after leaving DR ($p = 0.001$)	High Frequency Ventilation after leaving DR ($p < 0.001$)
NEC surgery ($p < 0.001$)	NEC surgery ($p = 0.04$)
RDS ($p < 0.001$)	RDS ($p < 0.001$)
Pneumothorax ($p < 0.001$)	Pneumothorax ($p < 0.001$)
Focal Gastrointestinal Perforation ($p < 0.001$)	—
PIVH grade 3-4 ($p < 0.001$)	PIVH grade 3-4 ($p < 0.001$)
—	Cystic Periventricular Leukomalacia ($p < 0.001$)
Early Bacterial sepsis and/or meningitis (before day 3) ($p < 0.001$)	Early Bacterial sepsis and/or meningitis (before day 3) ($p = 0.078$)
Major Birth Defect ($p = 0.08$)	Major Birth Defect ($p = 0.025$)
PDA Ligation ($p < 0.001$)	—
Indomethacin/Ibuprofen use ($p < 0.001$)	Indomethacin/Ibuprofen use ($p < 0.001$)

again tested (Dehejia and Wahba, 2002; Pattanayak et al., 2011). We proceeded using this strategy in a systematic way in an attempt to achieve proper balance (for both PS and selected covariates) in all blocks as already described.

Results

Table 1 shows recorded pre-treatment variables and their association to treatment use and/or the outcome of interest. Seventeen variables were identified that behaved as true confounders (associated to treatment decision and outcome of interest). One additional variable (Cystic Periventricular Leukomalacia) showed a strong association with death in the first 28 days but was not related to use of surfactant. Those variables were included in the initial, full model to estimate the PS. Four additional variables that showed only significant association with treatment choice were not included (Brookhart et al., 2006; Austin, 2011). The final model retained 13 variables (Table 2).

Table 2: *Variables included in the final model.*

Mode of delivery
Gender
Endotracheal intubation in DR
Adrenaline /Epinephrine in DR
Prenatal corticosteroid use
Conventional Ventilation after leaving DR
High Frequency Ventilation after leaving DR
NEC surgery
RDS
Pneumothorax
PIVH grade 3-4
Indomethacin/Ibuprofen (therapeutic)
Gestational Age

Table 3: *Minimum and maximum values of estimated PS for treated and untreated groups.*

	min	**max**
Untreated	0.0034	0.9970
Treated	0.0059	0.9997

Summary statistics and distributional graphics (not shown) warned about the existence of lack of overlapping between the groups in both tails of the distribution of estimated PS. Table 3 shows minimum and maximum values of PS for both groups. As a consequence 102 untreated neonates in the lower tail and 104 treated neonates in the upper tail were dropped out of the analysis to obtain estimates in the common support region. This area therefore consisted of 4,133 newborns, out of which 1,971 were given surfactant.

A graphical display of the estimated density functions of PS for treated and untreated individuals illustrated the fact that, even after trimming the tails, the overall degree of overlapping was rather small over the whole range of observed PS values. Most treated newborns had very high PS values whereas most untreated ones had very low PS values (Figure 2).

We then split up the PS in quintiles and evaluated the extent of PS balance between groups (Figure 3). Balance based on statistical significance was obtained only in the first and fifth quintiles. Further splitting up the middle quintiles did not correct for the lack of balance in these regions of the PS values. A last additional subdivision of blocks led to a final division in ten blocks which achieved balance at a significance level $\alpha = 0.01$ but still with apparent uneven distribution of groups (Figure 4).

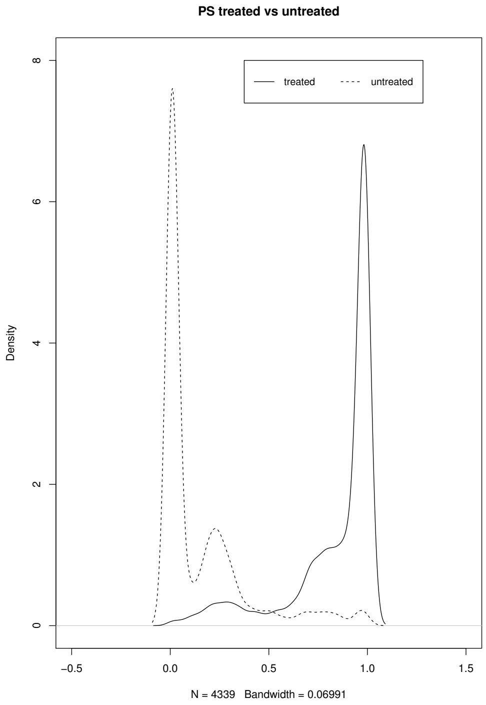

Figure 2: *Empirical distribution of PS for treated and untreated newborns (kernel function).*

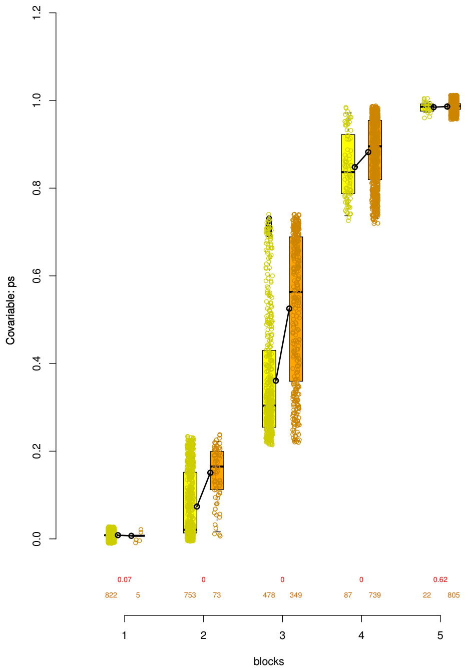

Figure 3: *Box-plots of quintiles of estimated PS. Second line over x-axis shows p values (Kolmogorov-Smirnov test of equivalence of distributions). First line, below, displays number of untreated and treated neonates respectively for each quintile.*

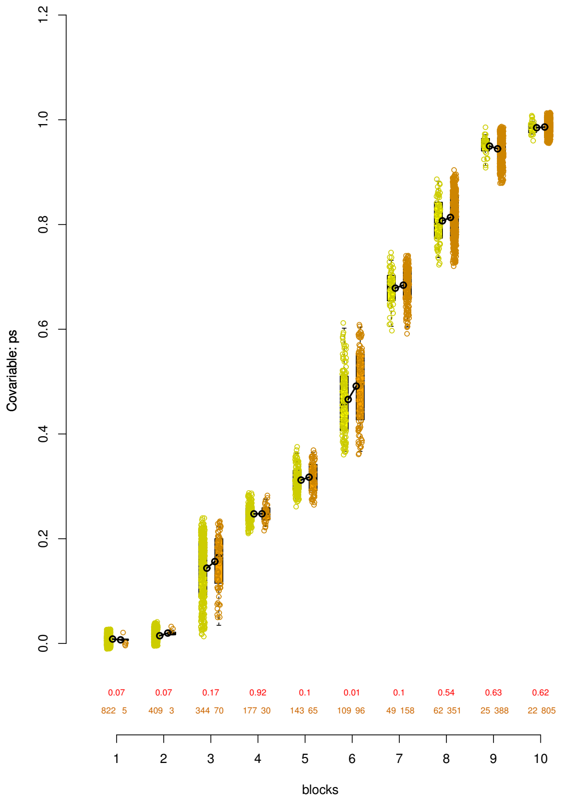

Figure 4: *Box-plots of estimated PS split-up in ten subgroups. Second line over x-axis shows p values (Kolmogorov-Smirnov test of equivalence of distributions). First line, below, displays number of untreated and treated neonates respectively for each subgroup (see text for further explanation).*

We went on to assess the ability of estimated PS to balance the distributions of individual baseline covariates between treated and untreated. Although reasonable balance seemed to be achieved for some variables, this was not so for some others. Figure 5 displays the comparative distribution of the variable “use of high frequency ventilation after leaving delivery room” showing the scarce number of treated neonates in the lower blocks and the lack of adequate balance in the sixth block ($p < 0.001$).

We decided not to proceed to the effects estimation stage as it was felt our estimated PS did not fulfill the theoretical assumptions required to provide unbiased, reliable adjusted estimates of the effect of surfactant administration on death during the first 28 days after birth.

Discussion

Given the frequent constraints to the conduct of randomized experiments in medicine, it is increasingly common to use observational data to assess the effects of clinical or

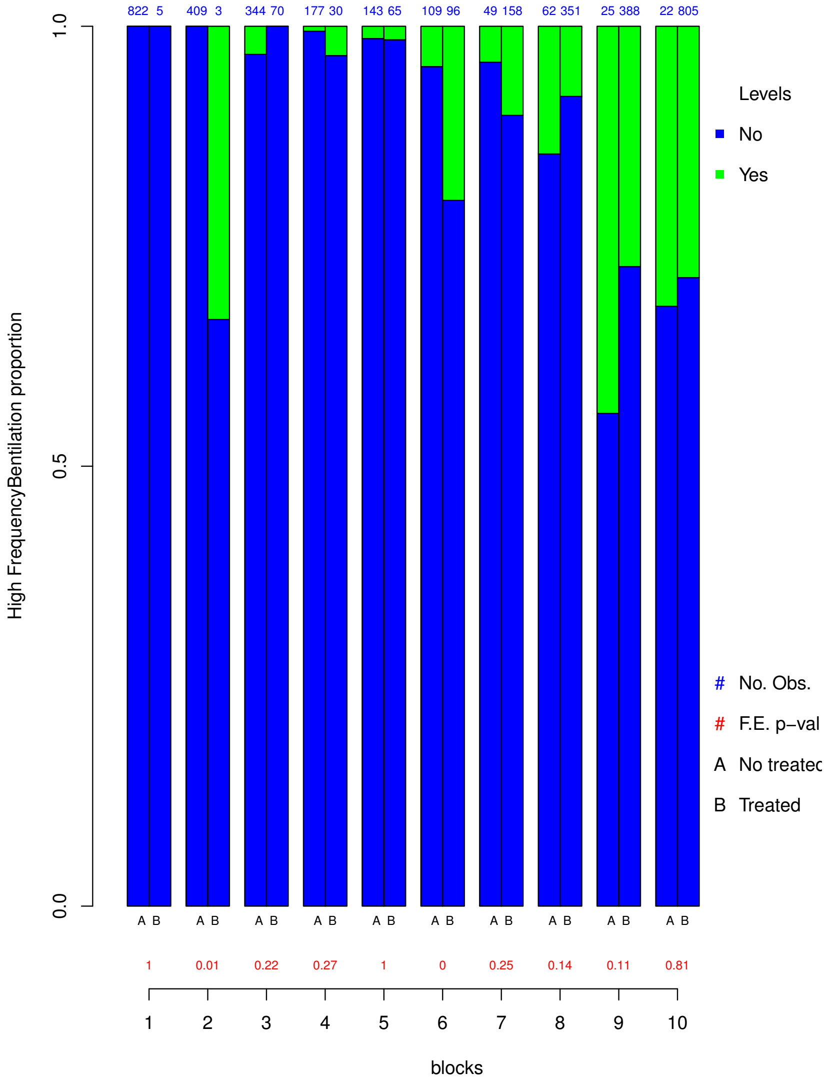

Figure 5: *Distribution of the variable "high frequency ventilation" showing the degree of balance achieved across the range of values of estimated PS. Above x-axis p values from exact Fisher test are shown for each block. Above each column, number of individuals is provided for untreated (columns A) and treated newborns (columns B).*

public health interventions on relevant health outcomes. In order to minimize the risk of obtaining biased estimates due to existence of unbalanced pre-treatment covariates in the groups to be compared (selection bias), an array of adjustment techniques has been devised (Gang et al., 2012). Among them those based on the PS have been gaining increasing popularity (Sturmer et al., 2006; Klungel et al., 2004). Nowadays there are routines in most popular statistical packages that make it easier to apply PS methods (Klungel et al., 2004; Becker and Ichino, 2002).

Our paper describes, using a case study, the full two-step process that should be undertaken to appropriately check for adequate conditional balance between groups. This process is commonly either not performed or not reported in published applications of these methods. It additionally shows that there may be specific settings where PS based adjustment should not be performed with the available observational data, as the estimated PS does not meet a critical theoretical requirement, namely being a balancing score. We comment on some of the undesired consequences of using PS based adjustment when this requirement is overlooked.

To obtain an empirical PS estimate that behaves as a true balancing score is a key step, dependent on several related factors that should be given proper attention. We would like first to emphasize the need for a correct selection of the covariates to be used in the PS estimation model. This process requires a careful combination of clinical knowledge on the issue at hand, a clearly framed causal pathway that takes into account all relevant information and a comprehensive consideration of the statistical relationships among pre-selected covariates, assignment of treatment and the outcome of interest (Brookhart et al., 2006; Bryson, Dorsett and Purdon, 2013).

One second related aspect is how to best determine that the estimated PS model will make for an appropriate adjustment. As the PS is an estimate of the probability that a given individual receives or not the study treatment and logistic models are commonly used to estimate the PS, it has been common practice to check the adequacy of the model employing standard goodness-of-fit (GOF) diagnostics. In particular, the c-statistic or area under the curve (AUC), an accepted measure of the ability of the model's predicted values to discriminate between positive and negative cases (Midi, Rana and Sarkar, 2010), has been reported in research employing PS adjustment methods (Westreich et al., 2011). High AUC values reflect good predictive performance of the estimated PS model, but the main concern in our setting is not to predict treatment selection but to control for confounding. Theory says that, conditional on PS (as a balancing function of covariates), treatment assignment or choice can be thought of as a random process (conditional independence assumption). If the treatment selection model has an extremely high predictive value, as our case study exemplifies (AUC=0.96), it is difficult to accept this assumption is met. In this model one or more factors strongly determine whether the individual receives or not the intervention under study and therefore it is doubtful that individuals from both treatment groups are "comparable". In this sense, the yield of a high c-statistic in the treatment choice estimation model must raise further concern that poor overlap between treated and untreated patients is likely to be an issue. Therefore it is now recognized that the c-statistic should not be provided as an index supporting the quality of the model. Several GOF statistics and graphical tools have been proposed aimed at specifically checking the adequacy of the PS model as a balancing score (Austin, 2008a).

The degree of overlapping in PS distributions between treated and untreated patients, the third related element, greatly influences comparability and therefore the quality of inference about treatment effect. The positivity principle, one of the key assumptions for causal inference (Westreich and Cole, 2010; Cole and Hernán, 2008), requires the existence of both treated and untreated subjects at each level of all covariates under consideration. This should also be reflected by the existence of individuals from both treatment groups in all regions of the PS range. However, PS estimated from models with very high predictive abilities will often lead to rather little overlap between treated and untreated (Sturmer et al., 2006). This suggests an inability to make fair comparisons between treated and untreated subjects (Glynn, Schneeweiss and Stürmer, 2006). It has further been shown that there is no association between the value of c-statistic for a given PS model and its ability to balance prognostically important variables between treated and untreated subjects (Austin, Grootendorst and Anderson, 2007).

In our example lack of overlap is small in the tails of PS distribution and therefore it might be considered, at first glance, that overlap is not a big issue (Table 3). However, as Figure 2 shows, most treated patients have very high values of PS whereas most untreated newborns have very low values. This finding supports the notion that, based on our observed pre-treatment covariates, there is low "randomness" (uncertainty) as to whether the patient is to be prescribed surfactant.

Two natural negative consequences of this lack of "conditional randomness" and subsequent absence of appropriate overlap arise: on the one hand, the need to restrict the estimation of treatment effects to a fraction of the study sample where this overlap holds, which in turn influences generalizability of results. On the other hand, the lack of balance achieved by the PS which leads to biased and unreliable effect estimates (Westreich et al., 2011).

If lack of appropriate balance in the PS is found, variables included in the PS estimation model should be carefully reviewed. Detailed assessment of the mechanism relating each variable to treatment choice/assignment and the magnitude of statistical association may help identify baseline covariates that behave as proxies of treatment allocation. If variables of this type are identified, they must be removed from the estimation model and the whole PS building process should start again. Sometimes refinement of the functional form of the regression model estimating PS, including higher order and interaction terms, may help achieve balance (Austin, 2011; D'Agostino and D'Agostino, 2007). Different specifications of our PS model did not provided any real remedy. It may be the case, though, that we have to conclude that the available observational data do not meet the required assumption of randomness of treatment assignment conditioned on a set of observed baseline covariates. This is tantamount to saying that these data are not adequate to obtain valid and reliable estimates of treatment effects.

This is, as far as we know, the first paper that presents a real case study where the balancing property of PS is not achieved. In our case, it was finally agreed to by involved clinicians that frequently treatment assignment decisions were largely determined by implicit clinical decision rules based on general knowledge and routine practice. Accordingly, in order to obtain valid estimates of the effect of surfactant given to premature newborns with respiratory problems further selection of specific subgroups and clinical scenarios where uncertainty about the beneficial effect of this treatment holds true should be sought.

It is expected that current growth in the use of PS methods continues as availability of and access to electronic data is on the rise (Couper and Miller, 2008) and ease of use of menu-driven general statistical packages also increases. There remains debate, however, on a variety of aspects related to the use of propensity score adjusted estimates and further research is warranted if its performance is to be maximized. We can mention, among others, selection of the best approach to obtain estimated PS likely to achieve balance (Imai and Ratkovic, 2014) and choice of a specific set of goodness-of-fit tools aimed at assess extent and quality of covariate balance achieved for each different implementation of the method (Austin, 2008a; Austin, 2009; Belitser et al., 2011). Above all, what is of critical importance is to improve the quality and standards in describing methods used to obtain the empirical PS and to adjust for it, as several articles claim poor and incomplete reporting is common (Austin, 2008b; Shah et al., 2005).

Our case study highlights several important issues: a) the need for careful consideration of whether information contained in available observational data allows for a treatment effect estimation question to be adequately addressed either globally or for some specific subgroup(s) of patients (Austin et al., 2005); b) the requirement that researchers thoroughly verify that the estimated PS truly achieves balance in treatment groups across the range of PS values as well as across levels and categories of the selected pre-treatment covariates. By so doing, clinicians and researchers will ensure that appropriate data and analytical methods are being used to obtain valid answers to focused clinical and public health questions on the causal effect of interventions. These results should help guide clinical practice when a randomized experiment is not feasible.

Acknowledgements

We wish to thank the contribution of doctor Adolf Valls i Soler† in this project and the European Neonatal Network (EURONEONET) supported by grants from the Directorate-General for Health and Consumers (DGSANCO).

This research was supported by the Health Department of the Basque Government (201211008) and Departament of Education, Language Policy and Culture of the Basque Government (IT620-13).

References

Adelson, J. L. (2013). Educational Research with Real-World Data: Reducing Selection Bias with Propensity Scores. *Practical Assessment, Research & Evaluation*, 18(15). Available online: http://pareonline.net/getvn.asp?v=18&n=15

Austin, P. C. (2008a). Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology Drug Safety*, 17, 1202–1217.

Austin, P. C. (2008b). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 2037-2049.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28, 3083–3107.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46, 399–424.

Austin, P. C., Mamdani, M. M., Stukel T. A., Anderson G. M. and Tu J. V. (2005). The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in Medicine*, 24, 1563–1578.

Austin, P. C., Grootendorst, P. and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*, 26, 734–753.

Becker, S. O. and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *Stata Journal*, 2, 358–377.

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold R. H. H., de Boer, A. and Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology Drug Safety*, 20, 1115–1129.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Serie B*, 57, 289–300.

Brookhart, M. A., Schneeweis, S., Rothmann, K. J., Glynn, R. J., Avorn, J. and Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149–1156.

Bryson, A., Dorsett, R. and Purdon, S. (2002). The use of propensity score matching in the evaluation of active labour market policies. *Policy Studies Institute and National Centre for Social Research. Working paper number 4.* http://eprints.lse.ac.uk/4993/1/The_use_of_propensity_score_matching_in_the_evaluation_of_active_labour_market_policies.pdf Last accessed 13 September 2013.

Caliendo, M. and Kopeining, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 1, 31–72.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168, 65-.664.

Couper, M. P. and Miller P. V. (2008). Web survey methods-Introduction. *Public Opinion Quarterly*, 5, 831–835.

D'Agostino, J. R. Jr. and D'Agostino, R. B. Sr. (2007). Estimating treatment effects using observational data. *The Journal of American Medical Association*, 297, 314–316.

Dehejia, R. H. and Wahba, S. (2002). Propensity Score-Matching methods for nonexperimental causal studies. *The review of Economics and Statistics*, 84(1), 151–161.

Friedman, L. M., Furberg, C. D. and DeMets, D. L. (1998). *Fundamentals of Clinical Trials*. 3rd ed. Springer, New York.

Gang, F., Brooks, J. M. and Chrischilles, E. A. (2012). Apples and oranges? Interpretations of risk adjustment and instrumental variable estimates of intended treatment effects using observational data. *American Journal of Epidemiology*, 175, 60–65.

Glynn, R. J., Schneeweiss, S. and Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*, 98(3), 253–259.

Grimes, D. A. and Schulz, K. F. (2002). Bias and causal association in observational research. *Lancet*, 359, 248–252.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society*, 76, 243–263.

Klungel, O. H., Martens, E. P., Psaty, B. M., Grobbee, D. E., Sullivan, S. D., Stricker, B. H. Ch. et al. (2004). Methods to assess intended effects of drug treatment in observational studies are reviewed. *Journal of Clinical Epidemiology*, 57, 1223–1231.

Midi, H., Rana, S. and Sarkar, S. K. (2010). Binary response modelling and validation of its predictive ability. *WSEAS Transactions on Mathematics*, 9, 438–447.

Oakes, J. M., Johnson, P. J.(2006). Propensity score matching for social epidemiology. In: Oakes, J. M., Kaufman, J. S. editors. *Methods in Social Epidemiology*. John Wiley & Sons, San Francisco.

Pattanayak, C. W., Rubin, D. B. and Zell, R. (2011). Propensity score methods for creating covariate balance in observational studies. *Revista Española de Cardiología*, 64(10), 897–903.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and Rubin, D. R. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.

Shadish, W. R. and Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*, 10, 19–26.

Shah, D. R., Laupacis, A., Hux, J. E. and Austin, P. C. (2005). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, 25, 2084–2106.

Stuart, E.A. http://www.biostat.jhsph.edu/ estuart/propensityscoresoftware.html Last accessed 10 September 2013.

Sturmer, T., Joshi, M., Glynn, R .J., Avorn J., Rothman, K. J. and Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59, 437–447.

Walker, A. M. (1996). Confounding by indication. *Epidemiology*, 7, 335–336.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L. and Mor, V. (2004). Principles for modelling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13, 841–853.

Westreich, D. and Cole, S. R. (2010). Invited commentary: positivity in practice. *American Journal of Epidemiology*, 171, 678–681.

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A. and Sturmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety*, 20, 317–320.