Use of the propensity score matching method to reduce recruitment bias in observational studies: application to the estimation of drotrecogin alfa's impact on intensive care units workload.

Lionel Riou França, MSc, Stéphanie Payet, MSc, Katell Le Lay, MSc, Robert Launois, PhD. REES France, Paris, France.

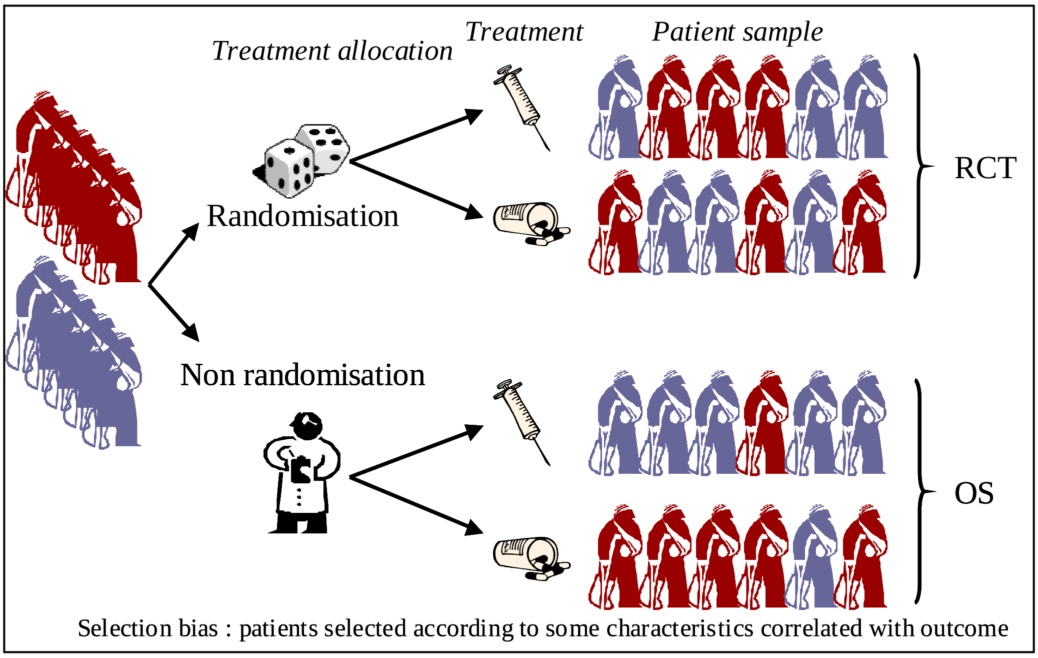
Based on a presentation given at the ISPOR 8th annual European meeting, Florence, Italy.

Context

Randomized clinical trials (RCTs) are considered the gold standard in clinical evaluations.[1] The main reason is that, when properly conducted, randomization ensures that treatment groups are comparable. Consequently, any difference detected is attributable to the intervention. As there is no need to control for confounding factors, the analyses are simpler. RCTs have good internal validity and are relevant for adoption decisions.
Sometimes, randomization is unfeasible, unethical or too costly. Moreover, non-randomized data may be already available. Observational studies (OSs) can then be an alternative to RCTs. They allow to measure the real-life practice and produce more generalisable results. Since these studies are expected to have a good external validity, they are relevant for policy decisions.

When treatment allocation is done according to the physician's decisions, we can expect some patients to be given preferentially one of the treatments, resulting in non-comparable groups. We are then in presence of recruitment bias (Figure 1). There is a need to correct for this bias when estimating a treatment's effect in a non randomized study.

Figure 1: Observational, non randomized studies do not guarantee comparability


Treatment allocation
Treatment
Patient sample
Randomisation
RCT
Non randomisation
OS
Selection bias : patients selected according to some characteristics correlated with outcome

The propensity score methods

The propensity score methodology allows to cope with the presence of recruitment bias.[2] The idea is to model, for each patient, the probability of receiving one of the treatments compared, according to a set of baseline characteristics. This figure is called the propensity score. The PS acts as a summary of all available information. If it is equally distributed among the patients of each treatment group, we can consider that the groups share the same characteristics.

Commonly, the PS is estimated using logistic regression. The presence of missing data among the baseline characteristics can therefore be troublesome. For instance, in the hypothetical case of 30 covariates independently missing for 3 % of the subjects, a listwise deletion of missing cases would lead to a reduction of 60 % in the sample size. To be able to estimate a PS for each patient, regardless of the presence of missing data, we used multiple imputation methods.

The criteria of a good regression model are well known in classical analysis: the model should be parsimonious and include only statistically significant predictors. The quality of the model can be quantitatively assessed using indicators such as Akaike's AIC. When modeling a propensity score, however, the issue is to ensure adequate balance in the patient's baseline characteristics. There is a need to include as much information as possible in the model.

In order to identify the most imbalanced covariates, we need a quantitative indicator. P-values are not the ideal candidate. Their value depends on the test selected (e.g. parametric or non parametric tests for quantitative variables) and on the sample size. Furthermore, the absence of statistical significance does not necessarily imply the absence of imbalance. A more appropriate summary statistic is the standardized difference (d) between treatment groups (Equation 1). This figure relates the difference in the groups' variable means to their observed variance.

Equation 1: The standardized difference statistic

$$d = \frac{\left(\overline{x_{treatment}} - \overline{x_{control}}\right)}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

We tested three logistic models using different variable selection strategies (Table 1). The first model strategy was the simplest: all measured baseline covariates were included in the model, without adding interaction terms. The second model selected only the most imbalanced or significant covariates. The third model used the same covariates as in the second model, but added the most significant interaction terms.

Table 1: PS Model selection strategies

Model	Strategy	Interaction terms
M1	Include all measured initial characteristics	No
M2	Stepwise selection of variables significant at the 10% level. Most imbalanced variables (d≥10%) forced in the model.	No
M3	Same as M2	Yes (10% significance level)

There are several ways to use the propensity score estimated. It could be used as an adjustment covariate, along with other outcome predictors. Alternatively, it could be used to weight the patients to make them representative of the population of interest. The PS can also be used to perform a stratified analysis. Finally, it can be used to match patients with similar propensity to receive treatment. The treatment groups in the matched sample are expected to share the same distribution of baseline characteristics, as in a randomized trial.

We chose to use propensity score matching since it leads to simpler analyses. We performed an optimal matching, where we tried to match each treated patient to a control minimizing the distance between the matched groups.

Three different matched samples were obtained, one for each propensity score model tested.

Application to the PREMISS study

Sepsis is a severe syndrome related to infection,[3] with high mortality rates. It is managed in France in Intensive Care Units (ICUs). Drotrecogin alfa has been shown to reduce mortality by 20% in the indication of severe sepsis,[4] and a medico-economic model lead to the conclusion that this new treatment was cost-effective in France, in the European treatment indication.[5]

The PREMISS study is an observational study carried out in France for the ministry of health to assess this new treatment's impact on healthcare. A control group was recruited before the drug's market authorization; the treatment group was recruited once the drug received its authorization. Eighty-eight intensive care units participated in this multicenter, pre/post study. Data was collected in a decentralised fashion, using an online case report form. In order to control for recruitment biases, forty-six baseline characteristics were retained.

Overall, 1096 patients were included in the study, 587 being in the treatment (i.e. drotrecogin alfa) group.
There is some evidence of recruitment bias in the study, since the control group tends to have smaller propensity scores than the treated group (Figure 2).
However, there is satisfying overlap in the groups' propensity scores, indicating that matching is conceivable.

Figure 2: Distribution of the PS among the treatment and the control groups (model M1)

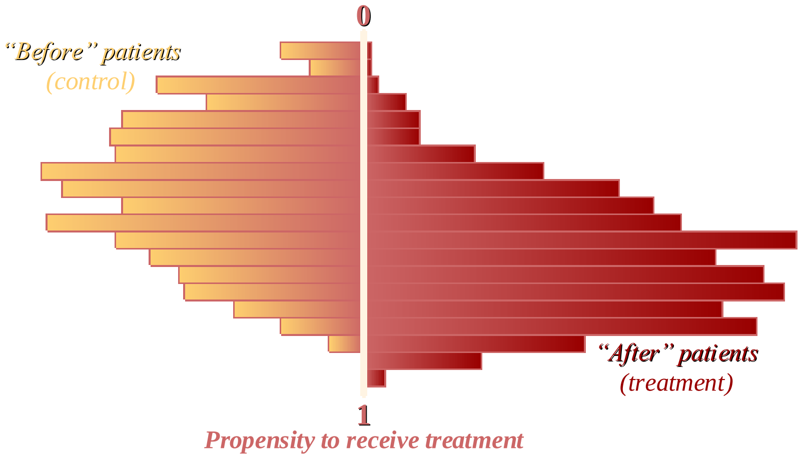


Table 2 summarizes the performance of the three PS models. In the resulting PS matched samples, model M2 keeps 79% of the patients and model M3 keeps 68% of them. The balance ratio is the ratio of the sum of the absolute values of the 46 standardized differences in the initial sample by the same sum in the matched sample. Model M1 performs best in reducing total imbalance, with a ratio of 2.39. However, some initial covariates remain unbalanced in all PS matched samples. Using a threshold of 10%, 2 baseline characteristics remain unbalanced in model M1, versus 5 in model M3.

Table 2: Performance of the three PS models in achieving a balanced matched sample

Criterion	Full Sample	Model M1	Model M2	Model M3
Sample size (%)	1096 (100%)	840 (77%)	870 (79%)	748 (68%)
Balance ratio[a]	1.00	2.39	2.08	2.16
Standardized differences d (%):				
Age class (≥80)	**25.51**	**14.98**	**14.31**	**15.56**
PaO_2/FiO_2[b]	**28.72**	**10.46**	8.03	4.42
Systolic blood pressure	9.35	6.92	**14.44**	6.95
McCabe score	**22.40**	8.59	**11.89**	6.45
LOD[c] renal subscore	**10.27**	9.17	**10.49**	**12.64**
Septic shock	8.95	8.21	5.72	**12.39**
Natremia	6.69	3.46	9.28	**11.39**
Admission category[d]	5.30	6.91	9.61	**10.37**

a: ratio = $\sum|d_{\text{full sample}}|/\sum|d_{\text{matched sample}}|$; b: "NA" in the absence of mechanical ventilation ; c: Logistic Organ Dysfunction score ; d: medical, scheduled surgery, emergent surgery, operated traumatism or unoperated traumatism

As the standardized differences indicate, patients included in the treatment group were younger and had less co-morbidities (as measured by the McCabe severity score). As age was entered in all PS models as a quantitative variable, neither of the PS models succeeded in achieving balance for the proportion of older patients.

Model M1 was selected for the remaining analyses, as it leads to the better balance between treatment groups.

One of the goals of the PREMISS study was to estimate drotrecogin alfa's economic impact on the intensive care units. The study collected a thesaurus of medical acts as defined in the new French common classification of medical acts, the CCAM. Each act is associated with a relative cost index, allowing for the estimation of the global ICU workload. Since this workload is highly skewed, we used a gamma regression model to estimate its increase among drotrecogin alfa treated patients. A random effects model was fitted in order to account for the clustering of the patients among the intensive care units.

Table 3: Estimation of the increase in ICU workload using drotrecogin alfa

	Full Sample	PS Matched Sample
Crude analysis	28%	18%
Adjusted analysis[a]	19%	14%

a: multivariate model including age, presence of ventilation, blood urea, admission by external transfer, presence of neurological infection and presence of urinary tract infection (all covariates statistically significant at the 5% level in the full sample model)

Table 3 gives the workload increase estimates among drotrecogin alfa treated patients using four different methods. Without taking into account the presence of recruitment bias, a full sample analysis estimates that treating patients with the new drug will increase workload by 28%. This figure is overestimated, since the patients included in the control group tended to be more severe. When adjusting for the presence of recruitment bias, this estimate lowers to 19% in the full sample, a figure similar to the estimate obtained in the crude analysis of the PS matched sample (18%). However, further adjustments in the PS matched sample reduce this estimate to a 14% increase.

Conclusion

The PS methodology has shown to be at least as good as multivariate adjustment methods. Its ease of use and of communication can make it appealing. However, conducting a good PS analysis requires careful consideration of the initial characteristics to measure (a large number of variables will increase the burden of data collection). If performing PS matching, there is a need for sufficient overlap between the groups, and the sample size may be increased to take into account that the more extreme patients will be excluded.

More essential is the fact that the PS methods only take into account observed variables. There is still a possibility for the presence of hidden bias. Furthermore, the PS methods allow to reduce recruitment bias, but not necessarily to eliminate it.

Finally, PS matching will reduce the study's external validity, since only a subset of the treated patients is used for the analysis.

In conclusion, the PS is a useful tool for the analysis of observational data, but, as any other tool, it has some limitations that need to be kept in mind.

References

[1] Dunn D, Babiker A, Hooker M, Darbyshire J. The dangers of inferring treatment effects from observational data: a case study in HIV infection. Control Clin Trials. 2002 Apr;23(2):106-10

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

[3] Bone RC, Balk RA, Cerra FB, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. Chest 1992; 101(6):1644-1655

[4] Bernard GR, Vincent JL, Laterre PF, et al. Recombinant human protein C Worldwide Evaluation in Severe Sepsis (PROWESS) study group. Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med. 2001; 344(10):699-709

[5] Riou França L, Launois R, Le Lay K, et al. Cost-effectiveness of drotrecogin alfa (activated) in the treatment of severe sepsis with multiple organ failure. Intl J Technol Assess Health Care. 2006;22(1). In press