Why to use propensity score in observational studies? Case study based on data from the Czech clinical database AHEAD 2006–09

Simona Littnerova\textsuperscript{a}\textsuperscript{*}, Jiri Jarkovsky\textsuperscript{a}, Jiri Parenica\textsuperscript{b,c,d}, Tomas Pavlik\textsuperscript{a}, Jindrich Spinar\textsuperscript{b,c,d}, Ladislav Dusek\textsuperscript{a}

\textsuperscript{a}Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic
\textsuperscript{b}Internal and Cardiology Department, Faculty Hospital Brno, Brno, Czech Republic
\textsuperscript{c}Faculty of Medicine, Masaryk University, Brno, Czech Republic
\textsuperscript{d}International Clinical Research Center—Department of Cardiovascular Disease, University Hospital St Anne’s, Brno, Czech Republic

**Abstract**
Randomized clinical trials represent the gold standard of the evidence based medicine research; nevertheless they may not always be feasible or ethical and the researchers have to rely on observational studies or research databases. However, obtaining reliable results from these studies requires the elimination of potential influence of confounding factors. Fortunately, several statistical methods capable of identifying and reducing the impact of confounding factors exist. One of them is the propensity score which has been frequently used in recent times to estimate relevant clinical effects adjusted for given confounders. This work aims to provide a concise and practical guide to propensity scores by means of an easily understandable case study. The case study is focused on gender differences in mortality rates of patients with acute heart failure in the Czech research database AHEAD (Acute Heart Failure Database).

© 2013 The Czech Society of Cardiology. Published by Elsevier Urban & Partner Sp.z.o.o. All rights reserved.

1. Introduction

Evidence based medicine (EBM) should be the principal methodical approach and keystone of the current clinical research and practice; it plays a key role in assessing strength of the evidence of benefits and risks of therapies as well as diagnostic tests \cite{1}. In order to obtain unbiased statistical results, factors responsible for the causal effects have to be distinguished from variables which show only a random correlation with the study endpoint \cite{2,3}.

Randomized clinical trials are generally considered as the gold standard for EBM. The initial randomization in such trials increases the chance that the effect of treatment on evaluated endpoint is given by treatment itself because the other influencing factors (such as age and gender structure, comorbidities, etc.) are randomly distributed between the compared groups of patients. Unfortunately, these studies...
may not always be feasible/ethical or applicable for real life situations in clinical practice [2].

In these situations, the observational studies can provide sufficient evidence; their advantage – as compared to the randomized clinical trials – is the ability to capture the real health care practice and to bring results which can be extrapolated on the population level. However, obtaining sufficiently unbiased results from these studies requires sophisticated statistical methods because the prerequisite of a random distribution of influential factors other than treatment itself is not always fulfilled and, therefore, the treatment effect is not independent of confounding factors [3,4].

Nowadays, several statistical methods working effectively with multiple mutually correlated variables exist. One of these methods is the propensity score method introduced in 1983 by Rosenbaum and Rubin, who defined the score as a “conditional probability of assignment to a particular treatment on a given vector of observed covariates” [5]. In other words, the score represents the patients’ probability of belonging to the same comparable population based on their descriptive characteristics, i.e. confounding factors. The main aim of this method is to obtain an unbiased estimate of treatment effect adjusted for the impact of given confounding factors in non-randomized and observational studies [3,4,6].

Due to its properties, the propensity score has recently become a widely used method in a broad range of epidemiological studies; its usage increased from 294 publications in the years 1998–2002 to 1111 in 2003–2007 and 3539 in 2008–2012.

Although several papers describe the methodology of the propensity score in detail [4,6], provide an example analysis on simulated data [7] or a systematic literature review [8,9], only few give examples of the application of propensity score on real data. The aim of this work is to explain the principles and advantages of the propensity score for clinical researchers. In addition to the theoretical background of the propensity score calculation, an analysis of influence of gender differences on mortality in acute heart failure patients is used as an educational case study (based on the Acute Heart Failure Database—AHEAD 2006–09 [10,11]).

2. Propensity score

The advantage of propensity score comparison in comparison to multivariable adjustment is the separation of confounding factors adjustment and analysis of the treatment effect steps [6]. Rosenbaum and Rubin [5] defined the propensity score for a patient as the conditional probability of being treated (exposed) given the vector of observed covariates. Let us assume that observed covariates \( X = (X_1, X_2, ..., X_n) \) and an indicator of treatment group \( Z = 1 \) if treated and \( Z = 0 \) if control) are assigned to each patient. The propensity score, denoted here as \( PS \), is defined for each patient as the probability that she/he received the treatment \( (Z = 1) \), given his/her observed covariates \( PS = P(Z = 1 | X) \). The probability in each patient of being either treated or untreated ranges from 0 to 1, i.e. \( 0 < PS < 1 \). The theory of propensity score relies on two major assumptions: (1) the observed variables do not affect the clinician’s decision that a patient will be treated (for example, gender/age of patients does not influence the administration of treatment) and (2) there are no unmeasured confounders (all the covariates potentially related to treatment assignment are known).

In an observational study, covariates are usually not balanced between treatment groups. Rosenbaum and Rubin [5] have demonstrated that observed covariates are balanced at each value of propensity score; it means that patients in treated and control groups with equal propensity score have the same distributions of the observed covariates.

The propensity score can be understood as a proxy between cases and covariates influencing the exposure, so it can be used instead of additional analyses of the covariates to simplify the analysis. Therefore, the propensity score as a proxy variable aggregates multiple confounding factors into a single dimension [4,5].

The simplest approach to estimate the propensity score is a simple tabular analysis where for each input variable it is obtained from a ratio of patients with and without treatment (Fig. 1); nevertheless, this approach is not suitable for continuous variables. To solve this problem, there are a plenty of other parametric and non-parametric methods for estimating the propensity score. For example, D’Agostino used the discrimination analysis to estimate the propensity score, while McCaffrey et al. published a paper where generalized boosted models were used to compute the propensity score [12]; Setouguchi et al. published a simulation study where neural networks and classification trees were adopted to estimate the propensity score. Nevertheless, these methods have major disadvantages such as their complexity and problems with subsequent interpretations, and their use is associated with considerable complications as a result of complex algorithms and their implementations [12]. For these reasons, the logistic regression has become the most frequently used technique to estimate the propensity score [4,9,13] and it was also used in our work.

Logistic regression is a statistical technique which estimates the probability of an event (the dependent variable) based on known factors (independent variables) which are expected to affect the occurrence of the event (treatment in the case of propensity score computation). The dependent variable can therefore only assume two values depending on whether the event (treatment) occurred or not. The result of logistic regression is the probability (ranging from 0 to 1) of an event occurring according to values of independent variables. Furthermore, logistic regression is a known and relatively easily understandable method for most researchers and it is implemented in all basic statistical software (SPSS, Stata, Statistica, SAS, R, etc.) [13].

The results of logistic regression can then be used to calculate the propensity score according to the following formula:

\[
PS = P(Z = 1 | X) = \frac{\exp (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}{1 + \exp (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}.
\]

where \( X_1, X_2 \) to \( X_n \) are independent covariates (in the example in Fig. 2a these are gender and age), \( \beta_0, \beta_1 \) to \( \beta_n \) are the corresponding regression coefficients, coefficient \( \beta_0 \)
represents the influence of an absolute component (the value of the probability when all covariates are equal to zero). Parameters $\beta_0, \beta_1$ to $\beta_n$ are estimated from the data using the maximum likelihood method \[14\] as shown in Fig. 2(a). After obtaining the propensity score, it is used to derive a pseudo-randomized dataset, allowing an unbiased estimation of the exposure (treatment) effect.

3. Application of propensity score for bias adjustment

The main and the most common approach of propensity score usage is prior to the quantification of treatment effect for data-matching; that is also described in our text. Other theoretical approaches (such as stratification and adjustment) were also developed but these are used less frequently \[4,5,9,13\].

Matching is a technique which creates a balanced dataset by making pairs from control and treated patients on the basis of a similar value of their propensity score; however, its major disadvantage is the fact that some patients may be discarded from the dataset. The matching method consists of selecting the first treated patient and finding the control patient with the same or nearly the same propensity score (Fig. 2b). To avoid matching of patients with largely different propensity scores, calliper is used to ensure that matching can only occur within a given range of propensity score values. This procedure may result in the situation when we have unpaired patients who are excluded from the subsequent analyses.

The efficacy of a propensity score model is estimated by the absolute standardized difference in individual covariates between groups of patients after matching. An absolute standardized difference 0% of a covariate indicates no residual bias for that covariate and an absolute standardized difference below 10% means an insignificant residual bias \[15\].

Matching based on the propensity score is frequently used in medical literature \[8,9,16\]. Several reasons contribute to the popularity of propensity score matching; matching can eliminate a greater portion of bias when estimating the more precise treatment effect as compared to other approaches \[17\]; matching by the propensity score creates a balanced dataset, allowing a simple and direct comparison of baseline covariates between treated and untreated patients. For these reasons, we have selected a case study of matching based on propensity score as a practical example of its most common usage. On the other hand, we should not forget that the balanced dataset obtained by matching does not contain all patients, and, therefore some information about the original dataset structure is lost.

3.1. Case study on the application of the propensity score in the acute heart failure mortality study (project AHEAD 2006–2009)

The following text explains how the propensity score works. In this example, we have chosen gender differences and their influence on mortality of patients. It is known that in population with heart failure men are younger than women and women have lower mortality rates than men \[18–21\]. The application of the propensity score allows us to obtain a balanced dataset and a more precise estimate of gender differences in mortality of patients (study endpoint). In this case study, gender represents the treatment indicator introduced in the theoretical part of this paper ($Z=1$ if male and $Z=0$ if female).

Data analyzed in this example came from the research database of acute heart failure AHEAD Main. The database AHEAD Main includes consecutive patients from seven centers with a 24-hour Catheterization Laboratory service and a centralized care for patients with acute coronary syndromes (ACS) from a region of about three million inhabitants \[10,11\]. The entire dataset contained 4153 patients, with 1757 (42.3%) women and 2396 (57.7%) men. In agreement with literature.
our results have confirmed that men with acute heart failure are younger than women with that condition (Fig. 3, Table 1). Moreover, men were also different from women in several other parameters (Table 1). The above-described propensity score methodology was applied on this primary dataset. Prior
to the computation of a propensity score using the logistic regression (as in example in Fig. 2a), we had to choose covariates which differ between genders and could represent confounding factors for the analyzed endpoint. The first step of variable selection consisted of the identification of important input variables both influencing mortality and Table 1 – Observed covariates before and after matching.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Women n=1757</th>
<th>Men n=2396</th>
<th>p*</th>
<th>Women n=1128</th>
<th>Men n=1128</th>
<th>p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>78 (55; 90)</td>
<td>69 (46; 86)</td>
<td>&lt;0.001</td>
<td>75 (52; 87)</td>
<td>75 (54; 87)</td>
<td>0.917</td>
</tr>
<tr>
<td>BMI</td>
<td>27.6 (20.3; 38.3)</td>
<td>27.7 (21.6; 37.8)</td>
<td>0.056</td>
<td>27.7 (20.5; 38.3)</td>
<td>27.7 (22.1; 37.5)</td>
<td>0.621</td>
</tr>
<tr>
<td>Systolic BP</td>
<td>140 (85; 210)</td>
<td>130 (80; 200)</td>
<td>&lt;0.001</td>
<td>140 (90; 210)</td>
<td>140 (90; 200)</td>
<td>0.050</td>
</tr>
<tr>
<td>Diastolic BP</td>
<td>80 (50; 110)</td>
<td>90 (55; 140)</td>
<td>0.154</td>
<td>90 (55; 145)</td>
<td>90 (55; 140)</td>
<td>0.209</td>
</tr>
<tr>
<td>Heart rate</td>
<td>89 (50; 145)</td>
<td>35 (15; 62)</td>
<td>&lt;0.001</td>
<td>40 (20; 65)</td>
<td>40 (18; 65)</td>
<td>0.367</td>
</tr>
<tr>
<td>Ejection fractionb</td>
<td>46.8 (18.6; 83.1)</td>
<td>54.2 (22.0; 91.8)</td>
<td>&lt;0.001</td>
<td>49.7 (21.9; 86.8)</td>
<td>50.5 (21.1; 84.8)</td>
<td>0.824</td>
</tr>
<tr>
<td>Creatinine clearanceb</td>
<td>46.8 (18.6; 83.1)</td>
<td>54.2 (22.0; 91.8)</td>
<td>&lt;0.001</td>
<td>49.7 (21.9; 86.8)</td>
<td>50.5 (21.1; 84.8)</td>
<td>0.824</td>
</tr>
</tbody>
</table>

AHF
De-novo | 1049 (60.9%) | 1320 (56.4%) | 0.004 | 683 (60.5%) | 558 (49.5%) | <0.001 |
ACS | 585 (33.3%) | 888 (37.1%) | 0.013 | 400 (35.5%) | 396 (35.1%) | 0.860 |

AHF syndromes
ADHF | 905 (51.5%) | 1335 (55.7%) | 0.008 | 620 (55.0%) | 668 (59.2%) | 0.051 |
Pulmonary edema | 330 (18.8%) | 418 (17.4%) | 0.265 | 220 (19.5%) | 200 (17.7%) | 0.279 |
Cardiogenic shock | 234 (13.3%) | 366 (15.3%) | 0.077 | 132 (11.7%) | 128 (11.3%) | 0.792 |

NYHA
I | 367 (20.9%) | 585 (24.4%) | <0.001 | 263 (23.3%) | 259 (23.0%) | 0.213 |
II | 569 (32.4%) | 629 (26.3%) | 375 (33.2%) | 316 (28.0%) | 0.279 |
III | 541 (30.8%) | 757 (31.6%) | 364 (32.3%) | 422 (37.4%) | 0.792 |
IV | 184 (10.5%) | 271 (11.3%) | 126 (11.2%) | 131 (11.6%) | 0.860 |

Medical history
Hypertension | 1306 (78.2%) | 1569 (69.3%) | <0.001 | 859 (76.4%) | 849 (75.3%) | 0.546 |
Diabetes mellitusb | 774 (46.2%) | 902 (39.8%) | <0.001 | 507 (44.9%) | 516 (45.7%) | 0.703 |
Myocardial infarctionb | 448 (26.7%) | 815 (36.0%) | <0.001 | 344 (30.5%) | 340 (30.1%) | 0.855 |
PCI and CABG | 7 (0.4%) | 14 (0.6%) | 0.399 | 5 (0.4%) | 7 (0.6%) | 0.562 |
Stroke or TIA | 302 (18.0%) | 351 (15.5%) | 0.038 | 186 (16.5%) | 189 (16.8%) | 0.865 |
Fibrillation or flutter | 489 (27.8%) | 593 (24.7%) | 0.025 | 299 (26.5%) | 337 (29.9%) | 0.075 |


a Statistical significance of difference between men and women, continuous parameters were tested by Mann–Whitney U test, categorical by ML Chi-square test.
b Parameters included in the propensity score model.
discriminating genders; it was based on the combination of results from data analysis, expert knowledge, and published results. All covariates shown in Table 1 were used in a forward stepwise selection to select covariates which were necessary to assemble the model. The final propensity score model consisted of four covariates: age, ejection fraction (EF), diabetes mellitus (DM) and myocardial infarction (MI) in medical history and creatinine clearance at admission (CC); the final regression equation was

\[ PS = \frac{P(Z = 1 | \text{Age}, \text{EF}, \text{DM}, \text{MI}, \text{CC})}{1 + \exp(3.94 - 0.04 \times \text{Age} - 0.03 \times \text{EF} - 0.18 \times \text{DM} + 0.47 \times \text{MI} + 0.01 \times \text{CC})} \]

The propensity score was then applied to match the structure of confounding factors for women with that for men, based on calliper 0.01 x standard deviation of the propensity score. The final balanced dataset consisted of 2256 patients (54.3% of original dataset): 1128 (50%) men and 1128 (50%) women. The decrease in the number of patients in the balanced dataset resulted in a shifted age structure; in total, patients in the balanced dataset were older than in the original dataset and the results obtained from the analyses of balanced dataset should be interpreted only for these older patients (Fig. 3).

The ability of propensity score to effectively balance the groups according to input variables was confirmed by combining the results in Table 1 and in Fig. 4. Table 1 shows that almost all covariates that entered the propensity score model were no longer significant in the matched sample. The absolute standardized differences of covariates between both genders after matching are shown in Fig. 4. The remaining covariates that significantly differed between genders (Table 1, Fig. 4) were due to the limited complexity of the propensity score model. Models with more covariates are better balanced but result in a smaller final dataset and so are less representative of the original dataset.

De-novo acute heart failure as the remaining significant covariate had only a limited influence on the outcome (its relationship with mortality was statistically non-significant). We have effectively balanced the dataset on age, DM, CC, EF, and previous MI in medical history. These factors differed between genders and are known as the most important in relation to mortality [10,11].

The main aim of our analysis was the comparison of men’s and women’s mortality rates. The difference in the mortality rates between men and women was not statistically significant in the original dataset (Fig. 5): the 5-year mortality rates were 58.0% for men and 60.4% for women (log-rank test \( p = 0.239 \)). In the balanced dataset, however, the 5-year mortality rates were 63.1% for men and 56.8% for women: this difference was statistically significant (\( p = 0.002 \)) (Fig. 6). The same pattern was found in the Norwegian Heart Failure Registry [20], where women had significantly lower mortality rates, too.

4. Conclusion

Propensity score is nowadays widely used in statistical analyses, particularly in the observational clinical studies. It is likely that its use will increase because the rising availability of data collection in research databases makes the observational design of studies more interesting and feasible [13]. The main advantage of the propensity score methodology is in its contribution to the more precise estimation of treatment response. Thus, the propensity score could be currently recommended as a standard tool for investigators trying to estimate the effects of treatments in studies where any potential bias may exist. Another promising application of the propensity score in health care studies is in the field of health technology assessment in comparing different health care facilities or health care providers. These units typically operate over a very different case-mix and the propensity score represents an effective approach to the extraction of mutually comparable samples from the real-life clinical data [22].

The use of propensity score matching in our case study revealed significant differences in mortality rates between both genders. This is due to the fact that the propensity score-matched dataset was balanced with respect to selected covariates; in other words, most of the bias influencing the mortality was removed using the propensity score.
methodology. Thus, we obtained more correct estimates of mortality rates for both genders.

Funding

This work was funded by grant projects of GACR no. 102/09/H083, of Masaryk University [MUNI/A/0828/2011], of the European Regional Development Fund—Project of International Clinical Research Center, St. Anne’s University Hospital Brno, Czech Republic (No. CZ.1.05/1.1.00/02.0123), and of the and by the project (Ministry of Health, Czech Republic) for conceptual development of research organization 65269705 (University Hospital Brno, Brno, Czech Republic).

Acknowledgment

We thank the AHEAD study investigators for their contribution to the study. Participating centers and investigators: Czech Republic: University Hospital Brno, Brno: Roman Miklik, Marian Felsoci, Katerina Horakova; University Hospital St. Anne’s, Brno: Jiri Vitovec, Lenka Spinarová, Tereza Mikusova, Klaudia Zidkova; Cardiocenter, University Hospital Kralovské Vinohrady, Prague: Petr Widimsky, Filip Rohac, Richard Fojt; General University Hospital in Prague, Prague: Ales Linhart, Jan Belohlavek; Na Homolce Hospital, Prague: Filip Malek, Petr Ošťádal, Andreas Kruger; University Hospital, Olomouc: Marian Fedorco, Jan Vaclavik, Cestmir Cihalik, David Vindis; Institute of Clinical and Experimental Medicine: Jiri Kettner; T.Bata Regional Hospital Zlin: Zdenek Coufal, Petr Hrdy, Stanislava Penasova, Miroslav Bambuch; Hospital Frydek-Mistek: Kamil Zeman.

REFERENCES


