Chapter 23
Comparative Effectiveness: Propensity Score Analysis

Kenneth P. Chen and Ari Moskowitz

Learning Objectives
Understand the incentives and disadvantages of using propensity score analysis for statistical modeling and causal inference in EHR-based research.

This case study introduces concepts that should improve understanding of the following:

1. Be aware of different approaches for estimating propensity scores: parametric, non-parametric, and machine learning approaches; and understand the pros and cons of each.
2. Learn different ways of using propensity scores to adjust for pre-treatment conditions, and to assess the balance of pre-treatment conditions among different treatment groups.
3. Appreciate concepts underlying propensity score analysis with EHRs including stratification, matching, and inverse probability weighting (including straight weight, stabilized weight, and doubly robust weighted regression).

23.1 Incentives for Using Propensity Score Analysis

When conducting research with electronic health records (EHRs) or other big data sources, we have access to a large number of covariates [1]. These covariates include patient demographics, physical parameters (e.g., vitals signs and physical examinations), laboratory parameters, home medications, pre-morbid conditions, etc. All these covariates could be confounders when considering the association between an exposure and an outcome. We can use statistical modeling to account for the confounding effect of these covariates and establish an association between the exposure and the outcome of interest [2, 3]. Propensity score analysis is

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_23

particularly advantageous when dealing with a large number of covariates [1]. The remainder of this chapter assumes a basic understanding of statistics and regression modeling (especially logistic regression).

Adjusting for as many covariates as possible sets the ground for a convincing causal inference by reducing latent biases due to latent variates [4]. However, this results in increased dimension [5]. Although large scale EHRs often have large enough sample size to allow high-dimensional study, dimension reduction is still useful for the following reasons: (i) to simplify the final model and make interpretation easier, (ii) to allow sensitivity analyses to explore higher order terms or interaction terms for those covariates that might have correlation or interaction with the outcome, and (iii) depending on the research question, the study cohort might still be small despite coming from a large database, and dimension reduction therefore becomes crucial for a model to be valid.

23.2 Concerns for Using Propensity Score

Although propensity score analysis has the above mentioned advantages, it is important to understand the theory of propensity score analysis and appreciate its limitations. A propensity score is an 'estimated probability' of one subject being assigned to either the treatment group or the control group given the subject's 'characteristics', or 'pre-treatment conditions'. It is a surrogate for all the covariates that are used to estimate it. It is not hard to imagine that using a single propensity score to represent all characteristics of a subject could introduce bias [6]. Therefore, implementing propensity scores in a statistical analysis model has to take into account the research question, the dataset, and the covariates included in the analysis. Furthermore, results must always be validated with sensitive analyses [7].

23.3 Different Approaches for Estimating Propensity Scores

In a randomized controlled trial, a causal relationship between exposure (treatment) and outcome can be readily determined if the randomization is carried out properly, i.e. if there is no difference in pre-treatment conditions between the two groups. However, in retrospective studies a difference in pre-treatment conditions between the two groups almost always exists. In order to demonstrate comparative effectiveness, causal inference with statistical modeling can be carried out in a number of ways [8, 9]. For propensity score analyses [3, 10], the pre-treatment conditions can be used as predictors in determining the likelihood of a subject being in the treatment group or the control group. In other words, the probability of being in the

treatment or control group is a function of pre-treatment conditions. There are a number of ways to generate this function. The most basic one is regression.

When using regression to estimate propensity scores, the outcome of the regression equation is either treatment group or control group, i.e. a binary outcome, and the variables in the regression equation can be a combination of numeric and nominal variables. This is a multivariate logistic regression that can be easily performed using most free or commercial statistical packages. If there is more than one treatment group (e.g., treatment A, treatment B, and control group) [11], then the propensity score can be estimated using a multivariate multinomial logistic regression.

The conventional regression model is a parametric model. Consequently, the estimated propensity score will be subject to any inherent limitations of the parametric model, i.e. model misspecification [12]. It is possible to use a non-parametric model to estimate the propensity score [13], such as regression trees, piecewise approaches, and kernel distributions. However, these methodologies are less established and are likely to require the use of machine learning algorithms [14]. Although non-parametric methods often require machine learning algorithms, machine learning techniques can be applied to both parametric and non-parametric methods. For example, some studies use a genetic algorithm to select variables and model specification for a conventional logistic regression to estimate propensity score [15].

23.4 Using Propensity Score to Adjust for Pre-treatment Conditions

The goal of using propensity score analysis is to create a treatment group and a control group that are indistinguishable from each other in terms of the pre-treatment conditions statistics (e.g., means and standard deviations of numeric variables, distribution of nominal variables). In other words, a treatment group and a control group are created that mimic a post-randomization assignment result of a randomized controlled trial, so that a causal inference can be made. Propensity score analysis is one of the tools to reach this goal [8, 9, 16].

For example, consider one subject that received the study drug or treatment (treatment group) and one subject that received placebo or standard treatment (control group). If they have similar pre-treatment conditions then their chance (probability) of being in the treatment group is the same. Consequently, it is comparable to two identical subjects being randomly assigned to either treatment or control group. When we find two subjects that have similar propensity scores where one actually received treatment and the other actually received placebo, we ‘match’ them in our final study cohorts before we look at the treatment effect (outcome variable). This process is called “propensity score matching.” By doing this, we will

have similar propensity score distributions (or pre-treatment conditions distributions) between the treatment and control groups.

If the model used to estimate propensity scores is well-specified [17, 18], we would expect the propensity scores to be representative of subjects' pre-treatment conditions. However, this might not always be the case, so we always look at the group statistics after propensity score matching. Since the ultimate goal is to eliminate the difference in pre-treatment conditions between groups, other methods like propensity score weighting have been proposed to achieve this. More sophisticated machine learning algorithms have also been developed that look at the balance of pre-treatment variables between two groups during the process of estimating a propensity score to ensure a valid model in simulating a randomized controlled trial-like result [19].

In EHR data research, we have access to a large number of pre-treatment covariates that we can extract from the database and use in the propensity score model. Although we cannot use an indefinite number of covariates to simulate a real RCT (which accounts for all unobserved variables), we can gain greater confidence in our conclusion by including more variables [20, 21]. Propensity score analysis is a powerful tool to simplify the final model while allowing a large number of pre-treatment conditions to be included. Figure 23.1 summarizes the above discussion of applying a propensity score model.

We now present a case study that used the MIMIC II database (v.2.26) [22, 23], and focus on the application of propensity scores in the analytic phase. The study was a retrospective cohort study of Intensive Care Unit (ICU) patients who were treated with at least one rate control agent (metroprolol, amiodarone or diltiazem). Propensity score analysis was performed using the following covariates: demographics, vital signs, basic metabolic panels, past medical conditions, disease severity scores, types of admission, and types of ICU. The outcomes measured

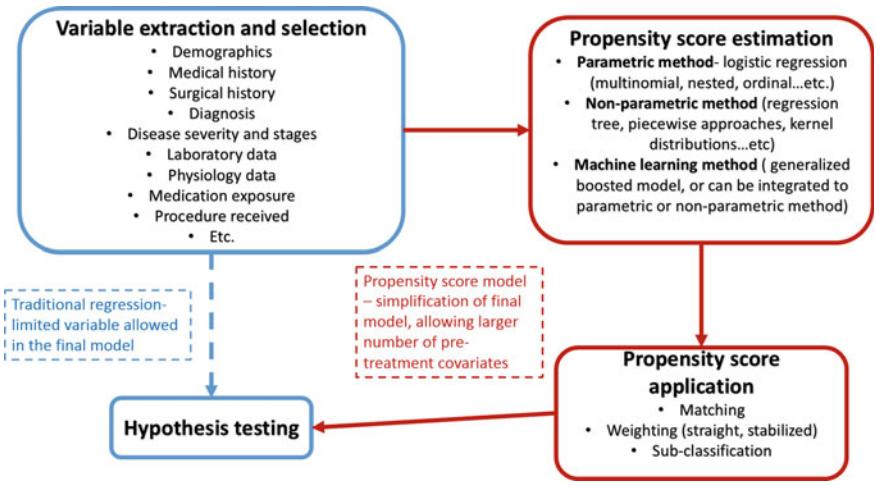

Fig. 23.1 Integration of propensity score analysis into a statistical design

were: (i) whether rate control was achieved by a single agent, or multiple agents (binary outcome); and, for those patients who reached rate control, (ii) the time to reach rate control (continuous outcome).

23.5 Study Pre-processing

In order to identify those patients with atrial fibrillation and rapid ventricular response (Afib with RVR) in the dataset, we used a combination of structured and unstructured data. Specifically, the structured data used included ICD-9 codes (the code for "Atrial Fibrillation" is 427.31) and medication administration data. The unstructured data used included waveform ECG data, serial heart rate (HR) data, discharge summaries and nursing notes. Unfortunately, only a small fraction of patients in the database have waveform data (approximately 2000 out of 32,000 patients). Consequently, we were unable to take full advantage of waveform analysis.

Patients who had Afib with RVR mentioned in their discharge summaries were identified by text searching equivalent keywords in discharge summaries while excluding the past medical history section. Once these patients had been identified we used the serial HR and medication administration data to find the subset of patients who had a HR of over 110 beats per minute (bpm) for more than 15 min and who received at least one of the rate control agents of interest (metoprolol, diltiazem, or amiodarone). Raw data was extracted using the Oracle® variety of SQL and was further processed using Python®, for text-searching discharge summaries, and Matlab®, for processing and plotting serial HR data and establishing temporal relationship between rapid ventricular response and medication administration.

Serial HR data existed for almost every patient in the database. However, contrary to the continuous waveform ECG data, it is only recorded every 5, 10, or 15 min and inconsistently. To make the data more homogenous and easier for plotting and processing, we interpolated the HR every 5 min: during the patient's ICU stay, if a raw HR data was not available for any given 5-min period, a value was interpolated using the two adjacent data points. Because of the infrequent sampling of HR for this data entity, one HR data point above 110 bpm would correspond to an episode of a rapid HR of 5-min duration. We arbitrarily chose a 15-min duration as a significant episode of rapid HR that warrants the algorithm (described below) to bring in more information from other data entity to determine if the tachycardic episode reflected Afib with RVR or another form of rapid rhythm (e.g. sinus tachycardia). This doesn't mean that a patient has to have 15 min of Afib with RVR before the physician decides to treat in clinical practice. Instead, it is a measure to reduce the noise of solitary rapid HRs. One can experiment on implementing different cut-off values and then review the result to determine an appropriate threshold.

After identifying an episode of rapid HR which appeared to last for at least 15 min, we next determined whether the patient received a pharmacologic control agent of interest within 2 h before or after the identified episode. A 2-h window was used because medication data and HR data are two different data entities, and the time stamps they carried might not be aligned exactly. Furthermore, the time stamps associated with medication data might subject to inaccurate data entry by human loggers. This window was arbitrarily determined; a smaller window would have increase specificity but decreased the sensitivity of detecting the cohort of interest, and vice versa for a larger window.

A major criterion for determining the effectiveness of a pharmacologic agent in the control of Afib with RVR is the time until termination of the RVR episode. As this information is not explicitly contained in the database, one has to define when the rate is 'controlled' and then run an algorithm to find the time lapse between the onset and resolution of RVR. The half-life of intravenous metoprolol and dilitazem are each approximately 4 h and, therefore, we defined the resolution of RVR as achieving sustained HR below 110 bpm for 4 h. Although there is no consensus for the definition of RVR resolution, as long as the same definition is used for every subject or sub-cohort, there is a ground for comparison. Our algorithm finds every HR below 110 bpm after the previous identified Afib RVR (episodes of rapid HR that lasted for at least 15 min and were treated by at least one rate control agent) and tested if the ensuing HR data in the following 4 h was below 110 bpm for at least 90 % of the time. The time lapse between the onset and the resolution can then be calculated.

Covariates, including demographics, vital signs, basic metabolic panels, past medical conditions, disease severity scores, types of admission, and types of ICU, were extracted using SQL. We also looked into the patient's home medication and past medical history of Afib. These pieces of information have to be extracted from the "home meds" and "past medical history" sections in the discharge summaries by using natural language processing techniques to text-search in a particular section of a discharge summary. Figure 23.2 is an example that our group used for discussing the analytic model.

Although we identified 1876 patients who were treated for Afib with RVR, only 320 of them received diltiazem as the first rate control agent. Using conventional regression analysis would result in over-fitting because of the small cohort size, and leaving out covariates would likely introduce biases. Propensity score analysis was used to reduce dimensionality. The first step is to estimate the propensity score (probability of being assigned to one treatment group given the pre-treatment covariates). As mentioned earlier, there are several different ways to estimate propensity scores including parametric methods such as multinomial logistic regression, and non-parametric methods such as prediction trees. Machine learning techniques can be implemented to train the propensity score model for optimized prediction. After the propensity score has been estimated, it can be used either as a variable in regression model to match subjects in different treatment groups with similar propensity scores, or to calculate inverse probability weights. When estimating propensity scores, besides optimizing the model to best predict the possible

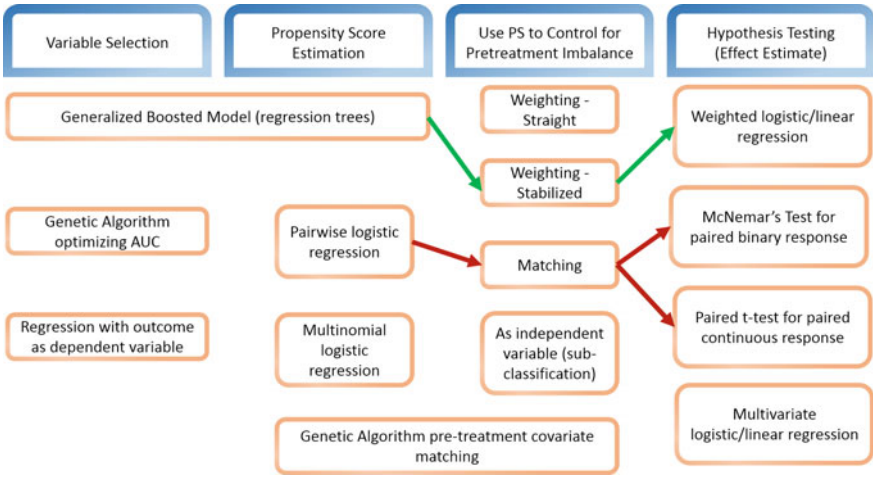

Fig. 23.2 Group discussions of the analytical model. The *green arrows* represent the final model, and the *red arrows* represent the model that was used as sensitivity analysis

treatment assignment given the pre-treatment variables, a newer concept is to estimate propensity scores to balance out pre-treatment covariates after matching or weighting. When using propensity score weighting, one can choose to use either straight weights or stabilized weights. Straight weighting is more susceptible to outliers with very distinct combination of pre-treatment covariates, and will double the cohort size when there are two treatment groups or triple the cohort size when there are three treatment groups. On the other hand, stabilized weighting is less susceptible to outliers, and does not increase the cohort size regardless of the number of treatment groups.

For this study we chose a machine learning algorithm (a generalized boosted model) to build a regression tree for the estimation of propensity scores (a non-parametric method). The reason for not choosing a parametric method is the same as that for not using a conventional regression analysis, as mentioned above. The model iteratively combines many simple regression trees until the pre-determined metrics for assessing between group pre-treatment covariate imbalance (standardized bias or Kolmogorov-Smirnov statistics) reach a minimum.

Extreme weights were eliminated using stabilized weights. Stabilized weights were then implemented in the final weighted regression for hypothesis testing. Depending on the nature of the outcome variable, weighted logistic regression is used for a binary outcome, and weighted liner regression is used for a continuous outcome. Several covariates with higher predictive power (of treatment assignment) were included in the final weighted regression model.

23.6 Study Analysis

In general, propensity score analysis has been used to compare two treatment groups, i.e. treatment versus control group. It is also commonly used for stratification (using propensity score as a covariate in a regression model) and propensity score matching (creating treatment and control groups of similar pre-treatment attribute and thus mimicking randomized trials). However, stratification can only establish association and propensity score matching mainly serves as a way of dimension reduction. Propensity score matching does carry the intention for causal inference, but matching propensity scores of three or more treatment groups requires calculating two or more dimensional distances for each matched group of subjects, which can be mathematically challenging and lacks supporting theory. Therefore, we chose machine-generated regression trees for our propensity score, and used a propensity score weighted regression model for outcome effect. The non-parametric approach avoided the limitations and biases introduced by model specification when using parametric methods. After the propensity score weight was generated, weighted regression was performed. This allows for exploration of interaction terms and adjustment for variables that have heavier effects on the outcomes that could not be fully eliminated by using propensity scores alone.

To validate our model, a series of sensitivity analyses using pair-wise propensity score matching were performed and similar effects of different treatment groups have on the outcomes were observed.

23.7 Study Results

In this single center retrospective cohort study, intravenous metoprolol was the most commonly used rate control agent for the control of Afib with RVR amongst patients in the intensive care unit. Using a novel propensity matching based approach, the effectiveness of metoprolol was compared to two other commonly used pharmacologic agents used for the control of Afib with RVR: diltiazem and amiodarone. With regards to the primary outcome of medication failure (defined as a switch to or addition of a second rate control agent), metoprolol had the lowest overall failure rate. Those patients who received diltiazem (odds ratio OR 1.55, confidence interval CI 1.05–2.3, $p = 0.027$) or amiodarone (OR 1.50, CI 1.1–2.0, $p = 0.006$) as their initial pharmacologic agent were more likely to receive an additional agent prior to the end of the RVR episode. In a secondary analysis of patients who received only one drug during their RVR episode, those who received diltiazem had significantly longer times to resolution of the RVR episode. Similarly, patients who received only diltiazem were also less likely to be controlled at 4 h than those who only received metoprolol (OR 0.59, CI 0.40–0.86, $p = 0.007$).

These results suggest that critically ill patients with Afib with RVR are less likely to require a second pharmacologic agent and more likely to be controlled at

4 h if they receive metoprolol as their initial rate control agent then either diltiazem or amiodarone. This effect seems to be most pronounced when comparing metoprolol to diltiazem.

23.8 Conclusions

While it is widely accepted that Afib with RVR in the ICU is associated with worse outcomes overall, there is no clear consensus with regards to optimal pharmacologic management and practice varies amongst clinicians. Through the use of a three-way propensity matching model, we have compared the most commonly used pharmacologic agents for this phenomenon and found evidence that starting with metoprolol may lead to fewer treatment failures and a more rapid resolution of the RVR episode.

Propensity score theory is more commonly implemented on two-treatment group studies. Estimating propensity score in multiple-treatment group studies and implementing that in causal inference can be statistically and mathematically challenging. In this chapter, we provided an example of multiple-treatment group propensity score analysis using machine-learning algorithm. The concepts explored in this chapter can be easily implemented in any two-treatment group studies. We also provided an example of two treatment group propensity score analysis in the sensitivity analyses of our study by performing pair-wise comparison between different treatment groups. Propensity score analysis can be a powerful way to achieve causal inference and dimension reduction in studies utilizing EHRs.

23.9 Next Steps

The data analysis strategy employed in this project may be particularly helpful in answering a range of research questions in the ICU setting. Critical care clinicians frequently have to select from a range of interventions or pharmacologic agents. As opposed to traditional propensity matching approaches where only two groups are compared, this model allows for the simultaneous comparison of three independent groups. Examples where this analysis approach could be useful include comparing the effectiveness of different vasopressors in the treatment of shock or different sedative agents for intubated patients with ARDS.

Given the degree of clinical equipoise with regards to the treatment of Afib with RVR in the ICU, the above results are powerful in providing some direction to clinicians faced with this complex clinical problem. Still, many questions remain. It is not clear, for instance, whether higher doses of diltiazem may have been more effective and thereby avoided relatively increased rates of treatment failure. We did not look at doses provided in this study. We also did not explore the oral versus intravenous versus combined routes of administration. Atrial fibrillation during

critical illness is a common phenomenon whose management requires further investigation.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The following key scripts were used:

- `database_query.sql`: used to extract data from the MIMIC II database.
- `data_extraction.m`: used to extract variables for analysis.
- `propensity_score_analysis.r`: used for propensity score analysis.
- `propensity_score_matching.r`: used for propensity score matching.

References

1. Patorno E et al (2014) Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. Epidemiology 25 (2):268–278
2. Fitzmaurice G (2006) Confounding: propensity score adjustment. Nutrition 22(11–12):1214–1216
3. Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 46(3):399–424
4. Li L et al (2011) Propensity score-based sensitivity analysis method for uncontrolled confounding. Am J Epidemiol 174(3):345–353
5. Toh S, Garcia Rodriguez LA, Hernan MA (2011) Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. Pharmacoepidemiol Drug Saf 20(8):849–857
6. Guertin JR et al (2015) Propensity score matching does not always remove confounding within an economic evaluation based on a non-randomized study. Value Health 18(7):A338
7. Girman CJ et al (2014) Assessing the impact of propensity score estimation and implementation on covariate balance and confounding control within and across important subgroups in comparative effectiveness research. Med Care 52(3):280–287

8. Glass TA et al (2013) Causal inference in public health. Annu Rev Public Health 34:61–75
9. Cousens S et al (2011) Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference. J Epidemiol Community Health 65 (7):576–581
10. Brookhart MA et al (2013) Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes 6(5):604–611
11. Feng P et al (2012) Generalized propensity score for estimating the average treatment effect of multiple treatments. Stat Med 31(7):681–697
12. Rosthoj S, Keiding N (2004) Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. Lifetime Data Anal 10 (4):461–472
13. Ertefaie A, Asgharian M, Stephens D (2014) Propensity score estimation in the presence of length-biased sampling: a nonparametric adjustment approach. Stat 3(1):83–94
14. Yoo C, Ramirez L, Liuzzi J (2014) Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J 18(2):50–57
15. Hsu DJ et al (2015) The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity score analysis. Chest 148(6):1470–1476
16. Hernan MA (2012) Beyond exchangeability: the other conditions for causal inference in medical research. Stat Methods Med Res 21(1):3–5
17. Austin PC, Stuart EA (2014) The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Stat Methods Med Res
18. Pirracchio R, Petersen ML, van der Laan M (2015) Improving propensity score estimators' robustness to model misspecification using super learner. Am J Epidemiol 181(2):108–119
19. Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. Stat Med 29(3):337–346
20. Brookhart MA et al (2006) Variable selection for propensity score models. Am J Epidemiol 163(12):1149–1156
21. Zhu Y et al (2015) Variable selection for propensity score estimation via balancing covariates. Epidemiology 26(2):e14–e15
22. Saeed M et al (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
23. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23):E215–E220