A review of propensity-score matching method in studies that exploring factors associated with COVID-severity

Introduction

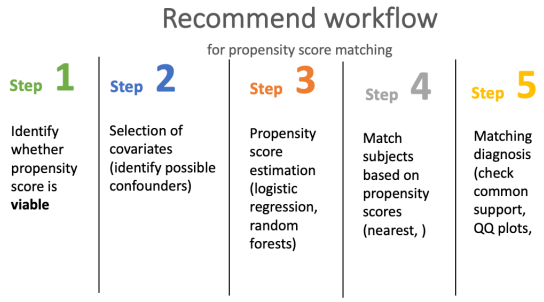
In the medical field, it is very common that the effect of a factor is of interest, such as therapeutic, preventive, adverse effects, etc. To be a well-designed study, usually, the outcomes, control, and the population of interest also need to be specified (PICO). Such a kind of factors are called treatments, exposures, or interventions depending on the context and purpose it was introduced. In this study, we didn't distinguish them and just used the term "treatments" to refer to them.

Randomized clinical trials (RCTs) are widely accepted as the golden standard in estimating the effect of a treatment. When perfectly implemented, subjects are randomly allocated into different treatment groups and balance is achieved both in measured baseline characteristics and unmeasured baseline characteristics. As a result, observed differences in outcomes between different treatment groups can be attributed purely to the difference in treatment. However, RCTs could be expensive to conduct since it artificially imposes a treatment whose effect is not yet fully understood to a group of subjects and requires a period of follow-up.

Instead, observational studies are commonly used as a substitution in exploring the potential effect of a treatment and RCTs can be introduced later if there is relatively strong evidence from observational studies. Unlike RCTs as an exemplar of experimental studies (or interventional studies), in an observational study, subjects are not assigned to different treatments at random (Cochran & Chambers, 1965). ~~One of the problems of observational studies is that~~. The treatment assignment is usually associated with baseline characteristics and among those confounders could exist. ~~People have developed different approaches to statistically adjust measured confounders in observational studies, such as regression models, analysis of covariance models, etc.~~ Propensity score matching (PSM) was one of the most popular methods used in the medical research field for adjusting confounders in observational studies. Propensity score, which is defined as the probability of receiving a treatment based on a subject's measured baseline characteristics (measured baseline covariates ~~as opposite to post-treatment covariates~~) can be used to match between the treatment and control group (ROSENBAUM & RUBIN, 1983). To estimate the propensity score of each subject based on their measured baseline covariates, a model that uses measured baseline covariates as predictors and treatment assignment as the outcome will be used, such a model is called the *propensity score model* to differentiate it from the analysis model that estimates the effect of treatment on the outcome of interest. After matching properly based on the propensity score, the measured baseline covariates will be balanced between the treatment and control groups. The propensity score matching method imitates the idea of RCTs. Theoretically, when measured baseline covariates used in the propensity score model included all the confounders, the confounders in data after matched should be balanced and the rest of the analysis can be done similarly as it is in the RCTs (it separates design and analysis).

A typical matching stage contains five steps. First, one needs to decide whether a propensity score matching method is viable based on the data available. It should be checked that whether the data has high-quality information about baseline covariates, treatment, and outcomes of interest. ~~It requires that the post-baseline characteristics (characteristics after treatment) should be similar. Later in this study, I will show cases in which this could be violated and sometimes strong assumption is needed to use the PSM method.~~ Second, select covariates used in the propensity score model. Third, select an appropriate propensity score model. Fourth, choose a right matching strategy to match subjects based on their propensity scores. Fifth, use appropriate balancing diagnosis methods to check whether balanced has been achieved between the treatment and control group after matching (Staffa & Zurakowski, 2018).

~~(Finally, one should stress that the propensity score model should only include variables that are measured at baseline and not post-baseline covariates that may be influenced or modified by the treatment, how about variables that different after treatment?)~~



Being relatively simple to use and easy to interpret the result makes it popular in the medical field. However, cautions in each step of using propensity score matching must be given or the result will be questionable. First, unlike RCTs will achieve balanced in either measured baseline characteristics and unmeasured baseline characteristics. Matching based on propensity score can only achieve a balance in those measured baseline characteristics. As a result, as many confounders as possible should be included in the propensity score model especially those that had been suggested in the published literature (Austin, 2011). Second, it is important to check the balance of measured baseline characteristics after matching to make sure that the propensity score model is correctly specified. Third, analysis methods that take the data dependence after matching into consideration are recommended (Austin, 2009; Imbens, 2004). Fourth, it should be aware that the treatment effect estimated by PSM is the average treatment effect in treated (ATT), which is the treatment effect on those who ultimately receive the treatment (Imbens, 2004). The reason why it was this specific treatment effect is estimated by PSM is determined by the underlying statistical theory and was out of scope in this study.

The emergence of massive electronic health record (EHR) databases makes it much easier to conduct observational studies. The capacity to extract more information than traditional observational studies also makes the propensity score matching even more powerful for the

reason aforementioned. Propensity score methods are also more suitable for the massive EHR than a multiple regression method because at least 10 events for every covariate entered into the regression model are suggested (Peduzzi et al., 1996) and in massive EHR data, the number of covariates extracted could easily reach to a large number. Besides, it could also benefit from the situation when the outcome is rare and treatment is more common (Braitman & Rosenbaum, 2002), ~~which is quite common in medical research~~. ~~Data analysts, data scientists, and domain experts who have basic statistical knowledge and want to use propensity score as a quick way of exploring the effect of a treatment while are not familiar with the detailed theory of propensity score matching could misuse it somehow. For example, they may feel lost in which covariates to be included in the propensity score model.~~ However, the complexity of EHR data may also make it hard to extract accurate baseline covariates and some baseline covariates cannot be extracted directly from the database, for which sophisticated computation or even combination with data from other sources is necessary. To make the best of massive EHR data, it is crucial to know which baseline covariates are more likely to be important, so researchers won't miss them and waste time in processing unnecessary baseline characteristics or just be satisfied with a large number of baseline characteristics in the propensity score model already.

Large-scale EHR-based observational studies are playing an important role especially in fighting the current COVID-19 pandemic. The effects of off-label treatments (treatments that were used beyond their original purpose) were of interest, such as preventing coronavirus infection and severe outcomes after infection. It could be unethical and sometimes almost impossible to randomly assign subjects to such kind of treatment and conduct RCTs to answer those questions (Harder et al., 2010). Large-scale EHR-based observational studies had shown their power in ruling out less promising treatments immediately after the outbreak of COVID-19 (Geleris et al., 2020) and provided evidence to push promising treatments forward to the stage of a RCT. Even for treatments whose effect for COVID has been studied in RCTs, the effect in populations that were not included in the original RCT design such as pregnant women or the elderly can be readily studied by large-scale EHR-based observational studies. But all those advantages must be based on the correct use of the propensity score matching technique.

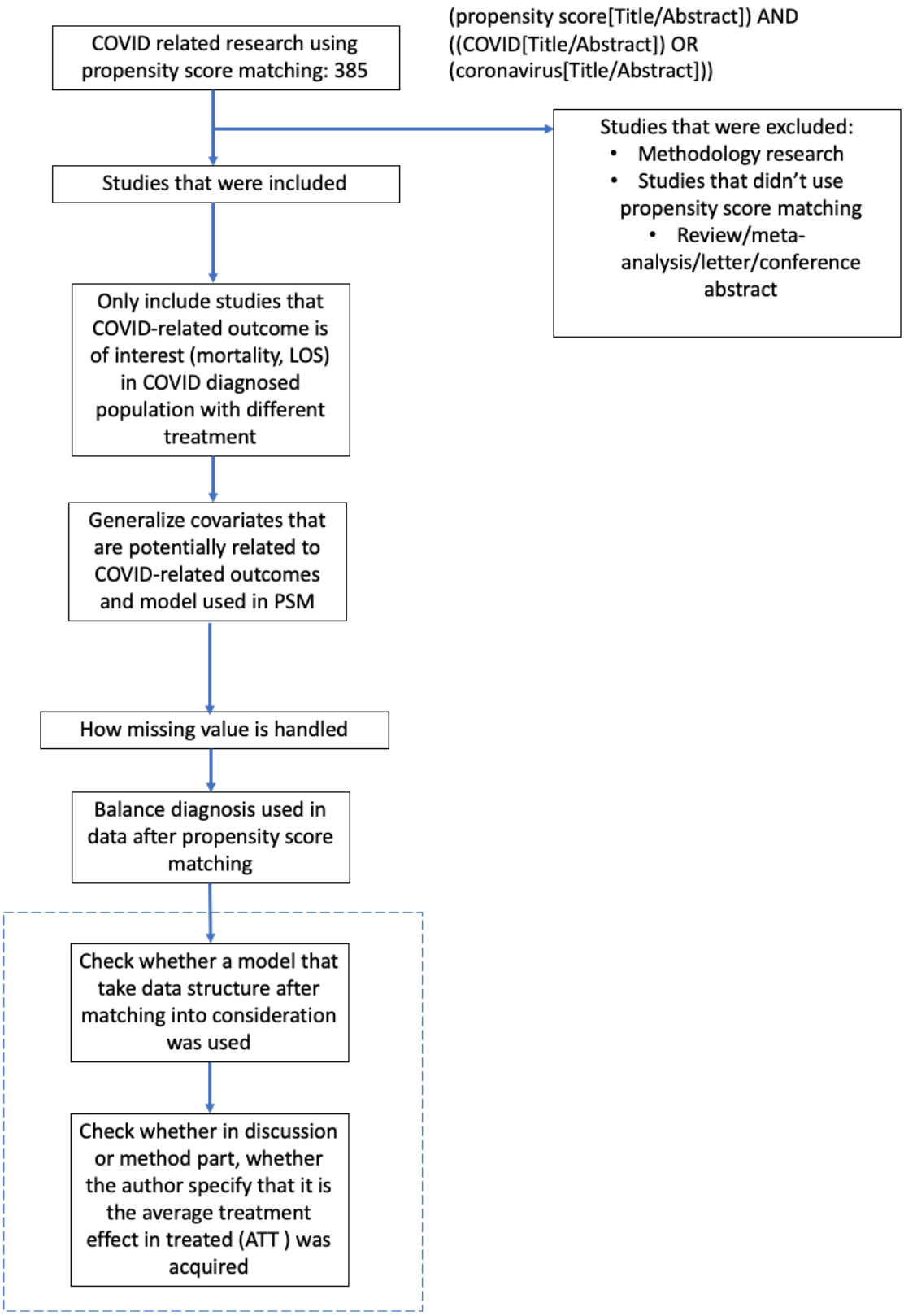
~~The variables selection step becomes more important in large-scale EHR data-based propensity score matching. Since now it is up to researchers to decide which variables to be extracted. The study will focus on the confounders and also check how each step of matching is done.~~

In this study, we will review a sample of studies on the COVID field that using propensity score matching as an approach to adjusting confounders. We will focus on studies that the effect of a treatment in COVID-related outcomes in a COVID positive population is of interest. We will check covariates used in those studies and try to generalize important potential confounders and functional form from them. The other important steps of using PSM that will also be reviewed to find strategies (which propensity score model to use, which matching strategy to use) in using PSM that could be optimal to this research topic from articles with high influence and discuss how the application of PSM in this area can be improved.

Method

Eligibility criteria

The method was according to PRISMA statement.


(propensity score[Title/Abstract]) AND ((COVID[Title/Abstract]) OR (coronavirus[Title/Abstract]))
COVID related research using propensity score matching: 385
Studies that were excluded:
• Methodology research
• Studies that didn't use propensity score matching
• Review/meta-analysis/letter/conference abstract
Studies that were included
Only include studies that COVID-related outcome is of interest (mortality, LOS) in COVID diagnosed population with different treatment
Generalize covariates that are potentially related to COVID-related outcomes and model used in PSM
How missing value is handled
Balance diagnosis used in data after propensity score matching
Check whether a model that take data structure after matching into consideration was used
Check whether in discussion or method part, whether the author specify that it is the average treatment effect in treated (ATT) was acquired

Search strategy

A search was conducted on 1st June on PubMed to identify COVID-related studies that using propensity score matching. We searched for articles with "COVID" in the title or abstract and "propensity score" in the title or abstract". The search was limited to publications between 2020 and 2021.

Selection process

A study was included if both of the two following criteria were satisfied: 1. The purpose of the study is to assess the therapeutic effect of a treatment (intervention, exposure) in COVID-related outcomes (mental outcomes were excluded) in the COVID diagnosed population. 2. Propensity score matching was used to adjusting possible confounders. Studies were excluded if they were reviews, methodological, letters, conference abstracts. Titles and abstracts were screened to assess eligibility. One reviewer was responsible for the whole selection process and no automation tools were used.

Data extraction

Eligible studies were checked manually by one reviewer to extract information regarding baseline covariates and their functional form used in the propensity score model, how missing values were handled, balance diagnostics. For eligible journals, the research areas and impact factors were obtained from the JCR website. The impact factors of journals were extracted from the 2020/2021 Journal Citation Report.

Data analysis

Studies included will be classified into two groups: high impact group and low impact group, based on impact factors of the journals where they were published and the impact factor of a journal in the year before the published year was used (for example, if an article was published in 2020, then the impact factor of the journal where it was published in 2019 was used).

The frequency table of baseline covariates and their most common functional form used in included studies will be given separately for each group and all included studies. A Chi-square test will be used to test whether the baseline covariates used in the two groups were different. Important baseline covariates will be generalized based on the frequency table. (The same comparison can be done between large-scale EHR based studies and traditional observational studies)

Similarly, frequency tables of choices of propensity model, methods of missing values handling, and methods of balance diagnostics used in the two groups will be created and chi-square tests will be used to test whether propensity score model, missing value handling, and balance diagnostics were different between the two groups separately.

Data items? (sample size, population, …)

Our data has limitation (info) but advantage too (EHR is not necessary continuous and we don't know even whether and where a record is missing compared with traditional obs study).

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*, *28*(25), 3083-3107. https://doi.org/10.1002/sim.3697

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*, *46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786

Braitman, L. E., & Rosenbaum, P. R. (2002). Rare Outcomes, Common Treatments: Analytic Strategies Using Propensity Scores. *Annals of Internal Medicine*, *137*(8), 693. https://doi.org/10.7326/0003-4819-137-8-200210150-00015

Cochran, W. G., & Chambers, S. P. (1965). The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society. Series A (General)*, *128*(2), 234. https://doi.org/10.2307/2344179

Geleris, J., Sun, Y., Platt, J., Zucker, J., Baldwin, M., Hripcsak, G., Labella, A., Manson, D. K., Kubin, C., Barr, R. G., Sobieszczyk, M. E., & Schluger, N. W. (2020). Observational Study of Hydroxychloroquine in Hospitalized Patients with Covid-19. *N Engl J Med*, *382*(25), 2411-2418. https://doi.org/10.1056/NEJMoa2012410

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234-249. https://doi.org/10.1037/a0019623

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, *86*(1), 4-29. https://doi.org/10.1162/003465304323023651

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*, *49*(12), 1373-1379. https://doi.org/10.1016/s0895-4356(96)00236-3

ROSENBAUM, P. R., & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Staffa, S. J., & Zurakowski, D. (2018). Five Steps to Successfully Implement and Evaluate Propensity Score Matching in Clinical Research Studies. *Anesth Analg*, *127*(4), 1066-1073. https://doi.org/10.1213/ANE.0000000000002787