Citation	<p>Moons P</p> <p>Propensity weighting: how to minimise comparative bias in non-randomised studies?</p> <p>Eur J Cardiovasc Nurs. 2020 Jan;19(1):83-88</p>
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	https://doi.org/10.1177/1474515119888972
Journal homepage	https://journals.sagepub.com/home/cnu
Author contact	<p>your email Philip.Moons@med.kuleuven.be</p> <p>your phone number + 32 (0)16 373315</p>
IR	

(article begins on next page)

Methods corner

Propensity weighting: How to minimize comparative bias in non-randomized studies?

Philip Moons ^{a,b,c}

^a KU Leuven Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium

^b Institute of Health and Care Sciences, University of Gothenburg, Gothenburg, Sweden

^c Department of Paediatrics and Child Health, University of Cape Town, Cape Town, South Africa

Word count: 2177

Address for correspondence: Philip Moons, KU Leuven Department of Public Health and Primary Care, Kapucijnenvoer 35, Box 7001, B-3000 Leuven, Belgium. Tel.: +32 16 373315.
E-mail address: Philip.Moons@kuleuven.be

Abstract

Non-randomized study designs are frequently used by researchers in cardiovascular nursing and allied professions. Baseline differences between the group to be compared may introduce biases in the results. Methods for causal inference address this issue. One such method is propensity weighting, in which two or more treatments/exposure groups are weighted to make the groups as comparable as possible. As such, it mimics a randomized controlled trial design. In this article, the Twang package is presented for propensity weighting, and its use is exemplified in a study on smoking and cannabis consumption in adults with congenital heart disease.

Learning objectives

- Recognizing the risk for bias in observational, non-randomized studies
- Knowing the role of propensity scores for causal inferences
- Understanding how propensity weighting can balance groups and yield more reliable findings on effects of treatments/exposures

The problem: Group comparison in non-randomized studies

In nursing and healthcare research, investigators predominantly use non-randomized study designs. Such observational studies aim to describe specific phenomena or patient populations, or they purport to compare groups of patients regarding particular characteristics. Also in cardiovascular nursing research, group comparisons based on data from observational, non-randomized studies are prevalent. Recent examples in the European Journal of Cardiovascular Nursing are studies where patients with or without sarcopenia after cardiac surgery (1); with or without social isolation in heart failure (2); with or without return of spontaneous circulation after cardiac arrest (3); or less than one year, one to three years and more than three years after heart transplantation (4) were compared.

In randomized studies, subjects are assigned to the groups at random. As a consequence, potential confounders are randomly distributed over the groups, which makes the groups comparable or balanced. Remaining differences between randomized groups are treated as a function of chance. Oppositely, in non-randomized studies, it is unsure if confounders are distributed at random. Hence, the groups to be compared may be imbalanced, and thus incomparable due to baseline differences. For instance, in a cohort study where hospital readmissions rates were compared between patients attending cardiac rehabilitation and those who did not, it was found that participation in cardiac rehabilitation was associated with less readmission (5). However, the study also showed that people with symptoms of unstable angina were less likely to attend cardiac rehabilitation than people without such symptoms (5). These baseline differences in unstable angina between the two group may be responsible for the differences in readmission rates, if not accounted for. Indeed, people with unstable angina may be readmitted more often than those without symptoms of angina, and thus bias the findings.

The fact that patients, and thus confounders, are not randomly distributed over the groups prevent researchers from drawing causal conclusion. Indeed, drawing conclusions in terms of ‘associations’ is the best that researchers can do in non-randomized studies. Sometimes this even leads to spurious correlations; which are associations found between factors that have nothing to do with each other. One funny example amongst many others is the spurious correlation of 0.95 between cheese consumption per capita and the number of people died by becoming tangled in bedsheets (www.tylervigen.com). In such an absurd example, the spuriousness of the correlation is obvious. However, in the scientific sphere, there are many spurious correlations that remain undetected, because authors and readers try to find logical explanations for the associations found.

Evaluating the effectiveness of healthcare interventions is a priority in research in cardiovascular care (6). Overall, randomized controlled trial (RCT) designs remain the gold standard to find causal effects. However, some research questions cannot be answered using a RCT, because of ethical or practical reasons. A well-known example is the impact of smoking on cardiovascular outcomes. Irrespective that a RCT is not possible, we want to come as close as possible to investigate causal effects. Methods for causal inference are therefore imperative.

A solution: Propensity weighting to reduce imbalances

A novel method of causal inference that aims at reducing imbalances between groups is propensity weighting. This technique is based on the calculation of propensity scores, which are individuals' probability of being assigned to the treatment/exposure group given observed baseline characteristics (7). Propensity scores allow researchers to create balance between treatment/exposure and comparison groups based on observed confounders (7). Indeed, propensity scores permit to separate the effect of a treatment/exposure on an outcome from observed confounding factors that influence both treatment assignments and outcomes (8), and thus mimics an RCT (7). This balance between the groups is not only created on one characteristic but on the full distribution of characteristics known of the groups. Since propensity scores are a probability, the values range from 0 to 1.

There are numerous ways to calculate propensity scores, such as logistic regression; Classification and Regression Trees (CART); or Generalized Boosted Regression Models (8). There are also different methods to utilize the obtained propensity score and remove or reduce the effects of confounding, such as stratification; matching; inverse probability of treatment weighting; or propensity weighting. Propensity score matching is probably the first and most widespread method. In propensity score matching, matched sets of treated and untreated

subjects who share a similar value of the propensity score are made (9). Recent examples in the European Journal of Cardiovascular Nursing where propensity score matching has been used are the cohort study mentioned above in which hospital readmission rates were compared between patients attending cardiac rehabilitation and those who did not (5), and a study on the effects of a standardized nursing intervention model on immobility in stroke survivors (10).

The main problem with propensity score matching is that researchers can seldomly use their entire sample, because not all subject of the control group have the characteristics to allow for the one-to-one paired propensity score matching without replacement (11). Alternatives have been developed in which the full study population can be used, and thus efficiency of the study is gained compared to traditional propensity-matching procedures (11). Propensity weighting is such a technique (8) and is an extension of inverse probability of treatment weighting that reweights both treatment groups to render the groups as comparable as possible in mimicking a RCT design (11).

Software

A user-friendly package to perform propensity weighting is Twang, which stands for 'Toolkit for Weighting and Analysis of Nonequivalent Groups' (8). Twang is developed by the RAND corporation and is a package that can be used in R, SAS, and STATA (8). It uses generalized boosted regression models to estimate propensity score weights. It is a form of machine learning that uses a non-parametric approach, checks all the potential confounders and interactions observed and entered in the model, handles missing data, and determines the optimal iteration based on balance (8). Doing so, the calculation of propensity scores is more comprehensive and stable than when using logistic regression, and given that it checks all the interactions between the covariates, it is truly extending what can be done by merely adjusting

for covariates in multiple regression. Twang provides various tools for assessing the quality of the propensity score weights and it handles both binary and more than 2 treatment group settings (8, 12, 13). RAND provides free tutorials and instructional videos on propensity weighting and the use of Twang: <https://www.rand.org/statistics/twang.html>

When using Twang, four steps need to be followed (8):

Step 1: Choose the primary effect of interest, being ‘average treatment effect in the population’ [ATE] or ‘average treatment effect in the treated’ [ATT] In ATE, the different groups are weighted. In ATT, only the control group is weighted.
(see tutorials for more information)

Step 2: Estimate propensity score weights

Step 3: Evaluate the quality of the propensity score weights in terms of convergence, balance and overlap

Step 4: Estimate the treatment/exposure effect on outcomes

Example of propensity weighting

A study where propensity weighting has been used had the aim to investigate the effects of smoking and cannabis use in adults with congenital heart disease (14). As part of the Assessment of Patterns of Patient-Reported Outcomes in Adults with Congenital Heart disease – International Study (APPROACH-IS) (15), health behaviors have been assessed in 4028 patients from 15 countries (16). To evaluate the effects of cigarette smoking and cannabis use on patient-reported outcomes, such as physical functioning, mental health, and quality of life (14). Cigarette smoking status and cannabis use were assessed by the Health-Behavior Scale–Congenital Heart Disease (17). Self-reported physical functioning was measured using the Physical Component Summary of the 12-item Short Form Health Survey (18). Mental health was assessed using the Mental Component Summary of the 12-item Short

Form Health Survey (18) and the anxiety and depression subscales of the Hospital Anxiety and Depression Scale (19). QOL was evaluated using a Linear Analog Scale (LAS) (20, 21) and the Satisfaction With Life Scale (22). The methodology of APPROACH-IS has been described extensively in a methods paper (23).

Patients were categorized into 4 groups: (i) patients who reported cigarette smoking only; (ii) patients who reported cannabis use only; (iii) patients with co-use of cigarettes and cannabis; and (iv) patients who did not use either cigarettes or cannabis. The group of non-users served as reference group. Given that the 4 groups of users differed according to baseline characteristics, the groups could not be directly compared without introducing bias. For instance, among smokers, 17% had a history of mood disorders, whereas this was only 12% in non-smokers. Hence, differences in outcomes could have been due to prior mood disorders rather than to smoking, if the groups would not be balanced. Therefore, propensity-weighting, and more specifically multinomial propensity-weighting with the four ‘treatment’ groups, was employed to balance the groups on available covariates that could potentially confound results of an unweighted analysis. Using Twang, the proposed four steps were followed. Step 1: An ‘average treatment effect in the population’ [ATE] was chosen, because we wanted to investigate what the effects are of smoking, cannabis and co-use in the population of adults with congenital heart disease. Step 2: The propensity score weights were estimated. The group assignment was modelled as the dependent variable and the demographic and clinical variables listed in Table 1 were included as covariates. For nominal variables, each response category is included as a separate covariate. Since APPROACH-IS included more than 4000 patients (23), we were able to adjust for this high number of covariates and did not need to make an a priori selection. If the sample were to be smaller, a process of a priori covariate selection would have been required. This could be done based on theory and prior research (24). Step 3: The quality of the propensity score weights in terms of

convergence, balance and overlap was evaluated. There were sufficient iterations to come to stable estimates (=convergence). Before propensity-weighting, 92 out 354 comparisons had standardized effect sizes greater than 0.2. After weighting, this was reduced to 34 comparisons (Figure 1). Further, on the Kolmogorov-Smirnov test, the p-value was below 0.05 in 34 of the 354 comparisons before weighting, and in 14 of the 354 comparisons after weighting. This means that the groups were more balanced after the weighting although small residual imbalances were observed. Finally, the 4 groups had sufficient overlap, hence the groups could be reliably compared. If there would have been little overlap, only subgroup analyses (on the overlapping subgroups) could have been conducted. Step 4: The exposure effect of smoking, cannabis and co-use on physical functioning, mental health, and QOL was estimated. Since there were remaining imbalances between groups (effect sizes >0.2), a linear model with doubly robust estimations was applied. ‘Doubly robust’ means that both the baseline covariates and the propensity weights were included in the regression model (8). It has previously been demonstrated that to obtain unbiased results, it is sufficient if one of the aforementioned models is correct (25).

PLEASE INSERT TABLE 1 AND FIGURE 1 ABOUT HERE

Propensity-weighting and doubly robust estimations showed that cigarette smoking had a significant negative effect on the physical component summary ($p=0.002$), mental component summary ($p<0.001$), anxiety symptoms ($p=0.033$), depression symptoms ($p=0.018$), LAS for QOL ($p=0.012$), and satisfaction with life ($p<0.001$) (14). For cannabis use, the only significant effect was on the physical component summary ($p=0.027$). Co-use had a significant negative effect on the mental component summary ($p<0.001$), anxiety symptoms ($p<0.001$), depression symptoms ($p<0.001$), LAS for QOL ($p=0.009$), and satisfaction with life ($p<0.001$) (14). For all the effects, effect sizes were computed, which are represented in Figure 2.

PLEASE INSERT FIGURE 2 ABOUT HERE

Because of the causal inference, the findings of this study go further than the typical descriptions of current health behaviors in patients with congenital heart disease (26). These results can guide nurses and other members of the interdisciplinary teams in their counseling efforts (27, 28), and the insights should be integrated in transitional care for young persons with congenital heart disease (29).

Reporting

When propensity weighting is conducted, it is important that information about the four steps is reported. More specifically, ATE or ATT as the primary effect (step 1); the variables used as covariates to calculate the propensity weights (step 2); and the indicators for convergence, balance and overlap (step 3) ought to be reported, before the real effect of the treatment/exposure on the outcomes can be described (step 4).

Conclusion

Group comparisons in observational, non-randomized studies may introduce bias since the groups often differ in baseline characteristics. Methods for causal inference are therefore imperative. Propensity weighting is a technique that is weighting two or more treatments/exposure groups to make the groups as comparable as possible. Doing so, it mimics a randomized controlled trial design. The Twang software is a user-friendly package that uses generalized boosted regression models to estimate propensity score weights. It is recommended that researchers in cardiovascular nursing and health use propensity weighting and similar techniques to avoid biases in their results.

References

1. Teng CH, Chen SY, Wei YC, Hsu RB, Chi NH, Wang SS, et al. Effects of sarcopenia on functional improvement over the first year after cardiac surgery: a cohort study. *Eur J Cardiovasc Nurs*. 2019;18(4):309-17.
2. Saito H, Kagiya N, Nagano N, Matsumoto K, Yoshioka K, Endo Y, et al. Social isolation is associated with 90-day rehospitalization due to heart failure. *Eur J Cardiovasc Nurs*. 2019;18(1):16-20.
3. Danielis M, Chittaro M, De Monte A, Trillo G, Duri D. A five-year retrospective study of out-of-hospital cardiac arrest in a north-east Italian urban area. *Eur J Cardiovasc Nurs*. 2019;18(1):67-74.
4. Wu YC, Tung HH, Wei J. Quality of life, demoralization syndrome and health-related lifestyle in cardiac transplant recipients - a longitudinal study in Taiwan. *Eur J Cardiovasc Nurs*. 2019;18(2):149-62.
5. Thomas E, Lotfaliany M, Grace SL, Oldenburg B, Taylor CB, Hare DL, et al. Effect of cardiac rehabilitation on 24-month all-cause hospital readmissions: A prospective cohort study. *Eur J Cardiovasc Nurs*. 2019;18(3):234-44.
6. Jaarsma T, Deaton C, Fitzsimmons D, Fridlund B, Hardig BM, Mahrer-Imhof R, et al. Research in cardiovascular care: a position statement of the Council on Cardiovascular Nursing and Allied Professionals of the European Society of Cardiology. *Eur J Cardiovasc Nurs*. 2014;13(1):9-21.
7. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424.
8. Ridgeway G, McCaffrey D, Morral A, Burgette L, Griffin BA. Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. Santa Monica, CA: Rand; 2017.

9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
10. Liu H, Zhu D, Cao J, Jiao J, Song B, Jin J, et al. The effects of a standardized nursing intervention model on immobile patients with stroke: a multicenter study in China. *Eur J Cardiovasc Nurs*. 2019:1474515119872850.
11. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*. 2013;9(2):215-34.
12. Burgette L, Griffin BA, McCaffrey D. Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package. Santa Monica, CA: RAND; 2017.
13. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*. 2013;32(19):3388-414.
14. Moons P, Luyckx K, Kovacs AH, Holbein CE, Thomet C, Budts W, et al. Prevalence and effects of cigarette smoking, cannabis consumption and co-use in adults with congenital heart disease from 15 countries. *Can J Cardiol*.
15. Moons P, Kovacs AH, Luyckx K, Thomet C, Budts W, Enomoto J, et al. Patient-reported outcomes in adults with congenital heart disease: Inter-country variation, standard of living and healthcare system factors. *Int J Cardiol*. 2018;251:34-41.
16. Holbein CE, Peugh J, Veldtman GR, Apers S, Luyckx K, Kovacs AH, et al. Health behaviours reported by adults with congenital heart disease across 15 countries. *Eur J Prev Cardiol*. 2019:2047487319876231.
17. Goossens E, Luyckx K, Mommen N, Gewillig M, Budts W, Zupancic N, et al. Health risk behaviors in adolescents and emerging adults with congenital heart disease: psychometric properties of the Health Behavior Scale-Congenital Heart Disease. *Eur J Cardiovasc Nurs*. 2013;12(6):544-57.

18. Ware JE, Kosinski M, Turner-Bowker DM, Sundaram M, Gandek B, Maruish ME. User's Manual for the SF-12v2 Health Survey Second Edition: QualityMetric, Incorporated; 2009.
19. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand*. 1983;67(6):361-70.
20. Moons P, Van Deyk K, De Bleser L, Marquet K, Raes E, De Geest S, et al. Quality of life and health status in adults with congenital heart disease: a direct comparison with healthy counterparts. *Eur J Cardiovasc Prev Rehabil*. 2006;13(3):407-13.
21. Moons P, Van Deyk K, Marquet K, De Bleser L, De Geest S, Budts W. Profile of adults with congenital heart disease having a good, moderate, or poor quality of life: a cluster analytic study. *Eur J Cardiovasc Nurs*. 2009;8(2):151-7.
22. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction with Life Scale. *J Pers Assess*. 1985;49(1):71-5.
23. Apers S, Kovacs AH, Luyckx K, Alday L, Berghammer M, Budts W, et al. Assessment of Patterns of Patient-Reported Outcomes in Adults with Congenital Heart disease - International Study (APPROACH-IS): rationale, design, and methods. *Int J Cardiol*. 2015;179:334-42.
24. Rubin DB. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodologies*. 2001;2(3-4):169-88.
25. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962-73.
26. Moons P, Van Deyk K, Dedroog D, Troost E, Budts W. Prevalence of cardiovascular risk factors in adults with congenital heart disease. *Eur J Cardiovasc Prev Rehabil*. 2006;13(4):612-6.

27. Moons P, De Geest S, Budts W. Comprehensive care for adults with congenital heart disease: expanding roles for nurses. *Eur J Cardiovasc Nurs.* 2002;1(1):23-8.
28. Thomet C, Moons P, Budts W, De Backer J, Chessa M, Diller G, et al. Staffing, activities, and infrastructure in 96 specialised adult congenital heart disease clinics in Europe. *Int J Cardiol.* 2019;292:100-5.
29. Moons P, Hilderson D, Van Deyk K. Implementation of transition programs can prevent another lost generation of patients with congenital heart disease. *Eur J Cardiovasc Nurs.* 2008;7(4):259-63.

Table 1. Demographic and clinical covariates included in the propensity-weighting

Age	History of arrhythmia
Female sex	History of mood disorders
Educational level - Less than high school	History of anxiety disorders
Educational level - High school	History of other psychiatric disorders
Educational level - College degree	Patient-reported New York Heart Association assessment - Class I
Educational level - University degree	Patient-reported New York Heart Association assessment - Class II
Employment status - Part-time or full-time work	Patient-reported New York Heart Association assessment - Class III
Employment status - Homemaker or retired	Patient-reported New York Heart Association assessment - Class IV
Employment status - Job seeking, unemployed, or disability	Country - Argentina
Employment status - Full-time student	Country - Australia
Employment status - Other	Country - Belgium
Marital status - Never married	Country - Canada
Marital status - Married or living with partner	Country - France
Marital status - Divorced or widowed	Country - India
Marital status - Other	Country - Italy
Complexity of the heart defect - Mild	Country - Japan
Complexity of the heart defect - Moderate	Country - Malta
Complexity of the heart defect - Complex	Country - Norway
Number of cardiac surgeries	Country - Sweden
Number of catheter interventions	Country - Switzerland
Congestive heart failure - No history of CHF	Country - Taiwan
Congestive heart failure - Past history of CHF	Country - the Netherlands
Congestive heart failure - Current CHF	Country - USA

Reproduced from Moons et al.(14) with permission of Elsevier....

Figure 1. Standardized differences between the treatment groups before and after weighting

Reproduced from Moons et al.(14) with permission of Elsevier

Figure 2. Effect of cigarette smoking, cannabis use, or co-use on patient-reported outcomes

Reproduced from Moons et al.(14) with permission of Elsevier