Propensity Score Methods in Rare Disease.
A Demonstration Using Observational Data in Systemic Lupus Erythematosus

Ibrahim Almaghlouth[1,3,4], Eleanor Pullenayegum[7,8], Dafna D Gladman [1,2,6], Murray B Urowitz [1,2,6], Sindhu R Johnson[1,5]

1 Division of Rheumatology, Department of Medicine, University of Toronto, Toronto, Canada

2 Centre for Prognosis in Rheumatic Diseases, University Health Network, Toronto, Canada

3 Rheumatology Unit, Department of Medicine, King Saud University, Saudi Arabia

4 College of Medicine Research Center, King Saud University, Saudi Arabia

5 Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

6 The Krembil Research Institute, University Health Network, Toronto, Canada.

7 Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

8 Program in Child Health Evaluative Sciences, SickKids Research Institute, Toronto, Ontario, Canada

Key Words. Propensity score, Observational, Selection bias, Balancing score

Word Count. 2013

Corresponding author. Sindhu Johnson MD PhD, Toronto Western Hospital, 399 Bathurst Street, Toronto, Ontario, Canada. Sindhu.Johnson@uhn.ca

Disclosures.
Dr. Johnson is a site investigator for Scleroderma clinical trials supported by Boehringer Ingelheim, Bayer, Corbus and GlaxoSmithKline.

Acknowledgements.
Dr. Johnson is supported by a Canadian Institutes of Health Research New Investigator Award.

Conflicts.
The authors have no conflicts of interest to disclose.

Abstract

Observational studies allow researchers to understand the natural history of rheumatic conditions, risk factors for disease development, factors affecting important disease-related outcomes and estimate treatment effect from real world data. However, this design carries a risk of confounding bias. A propensity score is a balancing score that aims to minimize the difference between study groups and consequently potential confounding effects. The score can be applied in one of four methods in observational research: matching, stratification, adjustment, and inverse probability weighting. Systemic lupus erythematosus (SLE) is a rare disease characterized by relatively small sample size and/or low event rates. In this article, we review propensity score methods. We demonstrate application of propensity score methods to achieve study group balance in a rare disease using an example of risk of infection in SLE patients with hypogammaglobulinemia.

Introduction

Clinical research in rheumatology can be complicated by the heterogeneity of many of the systemic autoimmune rheumatic diseases. Observational studies, such as case-control and cohort studies, provide a wider scope of patient representation, lower cost and longer follow-up time than traditional randomized trials. In addition, observational studies allow researchers to examine potential risk factors for clinically meaningful outcomes. However, these types of studies are criticized for the risk of confounding bias. Confounding of an exposure effect requires two features: association with the exposure of interest independently from the outcome, and independent association with the outcome but not on the causal pathway of the exposure to the outcome.(1-3) The presence of such a confounder is a threat to the estimated effect of the exposure. Small differences between groups in many variables can accumulate into substantial overall differences.(1) It may be that these differences have a greater effect on the outcome than the intervention itself.(4) This bias may result in a distortion of the measured treatment effect as a consequence of the way in which the study groups were constructed.(4)

In rheumatic disease research, investigators are also challenged by the rarity of the conditions. Small number of subjects are available for study. Furthermore, the numbers of events may be small. The small sample size can impact the ability to use conventional methodologic and statistical approaches to make inferences about treatment effects or risk estimates.(1, 5-9)

In this methodology article, we review propensity score methods as a potential solution to the risk of bias resulting from confounding, in particular when there are differences between the

exposed and non-exposed groups. Specifically, we demonstrate the applicability of propensity score methods in rheumatic disease studies with small sample size or low event rates, which are commonly encountered in the field of rheumatic diseases research. We highlight four propensity score methods. We discuss the use of standardized differences as a method to evaluate group differences before and after application of propensity score methods. We provide an example of how propensity score methods can be applied using observational data of a rare disease while comparing some of these commonly used methods. The aim of this article is to serve as a guide to clinical researchers, particularly in the field of rheumatology, who wish to apply propensity score methods.

Propensity score methods

A propensity score is a balancing score that can be used to account for the systematic differences between the exposure and control groups in an observational study.(10) The score is constructed by estimating the probability of exposure for each study cohort subject. This is achieved by conditioning probability of exposure on available observed variables. The propensity score can be estimated by regressing treatment assignment on observed baseline characteristics using a logistic regression model (Formula 1).(1) At an individual level, it is a measure of the likelihood that a person would have been treated considering their baseline characteristics.(1)

e(X) = P(Z=1|X)

Formula 1. Propensity Score

e(X) = propensity score
Z = exposure, where 1 = exposed, 0 = unexposed

X = a set of baseline characteristics, where X = ($X_{1....}X_p$)
P(Z=1|X) = probability of exposure given observed baseline characteristics
Note. Each patient has a probability of exposure where 0 < e(X) < 1.

The propensity score can then be used in one of four methods: matching, stratification, adjustment, and inverse probability weighting.

Matching. In this method, patients are matched based on their propensity score using a proximity method with predefined caliper width. This caliper width is based on the standard deviation of the logit of propensity score.(1, 11) Following that, adequacy of matching is assessed using either statistical testing or standardized difference between baseline covariates in the exposure and control groups.(12) The standardized difference is the absolute difference in the sample means divided by an estimate of the pooled standard deviation of the variable. The standardized difference represents the difference in means between the two groups in units of standard deviation. A similar formula is used for determining the standardized differences for dichotomous variables.(12, 13) (Formula 2 and 3).

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

Formula 2. Standardized difference for comparing means

d = standardized difference
$\bar{x}_{group}$ = mean of baseline characteristic in the specified group
s^2_{group} = variance of baseline characteristic in the specified group

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1-\hat{p}_{treatment}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}}$$

Formula 3. Standardized differences for comparing prevalences

d = standardized difference

$\hat{p}_{group}$ = prevalence of baseline characteristic in the specified group

There is some uncertainty on what constitutes an optimal standardized difference. Some authors use a standardized difference of 0.1 as upper limit of acceptable imbalance in baseline covariates, while others divide standardized difference into several cutoffs in which a difference less than 0.2 indicates low imbalance between matching groups, while 0.5 is moderate and 0.8 is considered large imbalance.(12) The utility of the propensity score was demonstrated by Johnson et al, where the investigators used matching on propensity score to improve group balance between patients with systemic sclerosis and pulmonary hypertension who were treated with warfarin compared to those who were not treated with warfarin.(7) The investigators demonstrated that group balance comparable to a randomized trial of similar size was achieved.

The matching method is best used when having a large pool of subjects in a control group. A significant loss of sample size may occur due to the lack of a match. In addition, this method will only account for variables that were included in the construction of the propensity score. Residual confounding may exist due to the impact of unmeasured confounders.(12, 14)

Stratification. Using this method, patients are stratified based on their propensity score. The exposure group and control groups are compared within each stratum. Wittenborg et al, applied this method to reduce confounding bias in a retrospective cohort study evaluating the

use of NSAIDs compared to an oral enzyme preparation, thought to have an effect on various rheumatic complaints.(15) Stratification based on propensity score may be limited by a reduction in sample size within each stratum, which may in turn reduce the power of the study to detect a treatment effect.(16) However, pooling across strata results in power reduction becoming less of an issue.

Adjustment. Using this method, the estimated propensity score is included in a regression model along with an indicator of exposure assignment. By doing this, within the context of a limited sample size, more confounding variables can be included to create the propensity score. The application of this technique was demonstrated in a study by Bergstra et al, in which the authors used propensity score adjustment when they compared the change in disease activity score in six or 12 months from the initiation of second treatment regimen of various disease modifying antirheumatic drugs (DMARDs) in rheumatoid arthritis patients who initially failed methotrexate. Patients were divided into categories based on the DMARDs that they received after the failure of methotrexate and the propensity score was used to adjust for confounding effect in the regression model.(17)

Inverse probability weighting. This unique method uses the propensity score to create a pseudo-population in which the exposure is unconfounded. This is achieved by weighting the exposure group based on the inverse of their estimated propensity score, while weighting the control group based on the inverse of 1-estimated propensity score. As a result, all subjects can be used in the study while reducing bias related to the systemic differences between exposure and control subjects (by giving appropriate weight based on estimated propensity score).(10) One of the caveats of this method is that it may lead to imprecise estimates if subjects have an extreme estimated propensity score (i.e. approximate to 0 or 1).(18) However, there are several proposed mechanisms to account for this occurrence, such as using stabilized weight.(19) Finally, the adequacy of balancing groups using this technique can be assessed by comparing the weighted average of the subjects' baseline covariates in both groups.(10, 12, 14, 16)

This method has been used by Kihara et al, to compare the effectiveness of tocilizumab to anti-TNFs when used as first biological therapy in rheumatoid arthritis patients using data from British biological registry. The authors in this study used propensity score inverse probability weighting to improve group imbalance between Tocilizumab and anti-TNF cohorts.(20)

Small sample sizes. Investigators often question how small the sample size can be to apply propensity score methods. Pirrachino et al reported a simulation study evaluating the impact of sample size on the performance of propensity score matching and inverse probability weighting methods. They found that reducing the sample size from 1000 to 40 subjects did not significantly impact the type 1 error rate. The inverse probability weighting method performed better than the propensity score matching method down to 60 subjects. When the sample size was 40, the propensity score matching estimators were either similarly or even less biased than the inverse probability weighting method estimators.(21)

Propensity scores in a rare disease. A demonstration.

Investigators interested in the use of propensity score methods often face the challenge of choosing which method to use. This may be particularly challenging in uncommon diseases such as systemic lupus erythematosus (SLE), where the number of subjects available for study may be limited due to rarity of the condition. SLE is a chronic autoimmune disease with a three-fold higher mortality than general population. Infection is a leading cause of death in this population. Defects in immunoglobulin synthesis or function could result in a significant risk of

serious infections. We aimed to assess whether acquired low levels of any type of immunoglobulin increases the risk of clinically relevant infection in adult patients with SLE. (22)

SLE patients in our long-term, single center, observational cohort were followed at 2-6 month intervals according to a standard protocol which included demographics, clinical, laboratory and therapeutic information.(22) Our study consisted of 437 SLE subjects with low immunoglobulins and 656 SLE subjects who never experienced low immunoglobulins and served as control subjects. The exposure (low immunoglobulin) was defined as the presence of two low immunoglobulin level measurements of the same type with the index date being the first measurement of low immunoglobulins. The primary outcome was clinically relevant infection defined as infection within two years of the index date requiring use of oral or parenteral antibiotics. The analysis was time to event using a Cox-regression model. There were 97 events, 47 in the exposure group and 50 in the control group. Patients with hypogammaglobulinemia had longer mean disease duration (11.2 ± 9.1 versus 7.6 ± 8.0 years), more frequently had a history of lupus nephritis (44.9% versus 17.8%), higher frequency of proteinuria (25.6% versus 11.3%) and more accumulated SLE damage (mean Systemic Lupus Damage Index score 1.2 ± 1.6 versus 0.5 ± 1.0).(22) (Table 1) Inability to account for these differences between groups would have led to biased estimation of the risk of infection.

We applied 3 propensity score methods to derive less biased estimates of the risk of infection in SLE patients with low immunoglobulins. We applied matching and inverse probability weighting propensity score methods separately to investigate our ability to achieve improvement of

group balance when comparing the risk of infection between SLE patients with and without acquired low immunoglobulins. We favored these two methods in particular because of previous studies that demonstrated minimal bias when used to estimate marginal effect.(23, 24) We also used propensity score adjustment due to its usability and the ability to retain the whole cohort. We did not use stratification on propensity score because of some criticism about its performance in reducing bias when dealing with few outcome events.(25) Variables used to construct the propensity score were age, sex, disease duration, disease activity measured by SLEDAI-2K score, nephrotic range proteinuria, antiphospholipid antibodies, prednisone use and dose, immunosuppressant use, and biologics use.(26) The choice of these covariates was based on the literature regarding associated or predisposing factors to low immunoglobulin states. The adequacy of balance was assessed using standardized differences.(12, 19)

Both propensity score matching and inverse probability weighting improved group balance. Table 1. Matching by propensity score demonstrated superior improvement in the standardized difference 8 of 11 (73%) of the variables. However, matching by propensity score resulted in smaller sample size (from 1093 subjects in the unmatched cohort to 922 subjects in the matched cohort) due to the loss of unmatched subjects. In comparison, inverse probability weighting was able to improve balance across all the variables. In addition, it allowed retention of the whole cohort (n=1093). Adjustment by propensity score was also applied and allowed for retention of the complete cohort. However, this method did not allow for evaluation of reduction in group imbalances.

Comparison of estimates of the risk of infection in SLE patients with and without low immunoglobulins using the three propensity score methods are presented in Table 2.
All three propensity score methods demonstrated that low IgA level significantly increased the risk of infection in SLE patients. Adjustment by propensity score had the greatest uncertainty around the estimate of risk (hazard ratio (HR) 3.19 (95% CI 1.17- 8.71)). Propensity score matching and propensity score weighting gave estimates of comparable magnitude and uncertainty, with propensity score weighting giving the most conservative estimate (HR 1.75, 95%CI 1.01, 3.02). Table 2.

This example illustrates the application of propensity score methods. It was previously believed that propensity score methods could only be used in large administrative databases. These methods are increasingly being successfully applied in observational data of rare diseases.(7, 27, 28). Furthermore, our study provides a comparison between several propensity score methods performances in reducing groups imbalance when applied to a survival model-based study with relatively small event rate. The robustness of the propensity score matching and inverse probability of treatment weighting methods in reducing potential bias due to measured confounders in our study was largely consistent with the simulation study by Pirracchio et al, in which the authors demonstrated good performance of both techniques when the sample size was as low as 40 subjects.(21)

Conclusion

In this paper we have described the use of propensity score methods to reduce the risk of bias in estimates of treatment effect or risk using observational data. We have highlighted their relative advantages and disadvantages. We have demonstrated the successful use of these methods in observational data of a rare disease, evaluating the risk of infection in SLE patients with low immunoglobulins. Rheumatic disease researchers may consider working with biostatisticians to apply propensity score methods to observational studies of rare rheumatic diseases.

References

1. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Propensity score methods for bias reduction in observational studies of treatment effect. Rheum Dis Clin N Am 2018;44:203-13.
2. Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis: A cautionary tale and an alternative procedure, illustrated by studies of british voting behaviour. Qual Quant 2018;52:1957-76.
3. VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat 2013;41:196-220.
4. Savitz DA. Interpreting epidemiologic evidence. Strategies for study design and analysis. Oxford: Oxford University Press, Inc.; 2003.
5. Johnson SR. Bayesian inference: Statistical gimmick or added value? J Rheumatol 2011;38:794-6.
6. Johnson SR, Feldman BM, Pope JE, Tomlinson GA. Shifting our thinking about uncommon disease trials: The case of methotrexate in scleroderma. J Rheumatol 2009;36:323-9.
7. Johnson SR, Granton JT, Tomlinson GA, Grosbein HA, Le T, Lee P, et al. Warfarin in systemic sclerosis-associated and idiopathic pulmonary arterial hypertension. A bayesian approach to evaluating treatment for uncommon disease. J Rheumatol 2012;39:276-85.
8. Johnson SR. Advanced epidemiologic methods for the study of rheumatic and musculoskeletal diseases. Rheum Dis Clin N Am 2018;44:xv-xvi.
9. Johnson SR, Tomlinson GA, Granton JT, Hawker GA, Feldman BM. Applied bayesian methods in the rheumatic diseases. Rheum Dis Clin N Am 2018;44:361-70.
10. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.
11. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A monte carlo study. Stat Med 2007;26:734-53.
12. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083-107.
13. Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. PharmacoepidemiolDrug Saf 2008;17:1218-25.
14. Austin PC. A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivar Behav Res 2011;46:119-51.
15. Wittenborg A, Bock PR, Hanisch J, Saller R, Schneider B. [comparative epidemiological study in patients with rheumatic diseases illustrated in a example of a treatment with non-steroidal anti- inflammatory drugs versus an oral enzyme combination preparation]. Arzneimittelforschung 2000;50:728-38.
16. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making 2009;29:661-77.

17. Bergstra SA, Winchow LL, Murphy E, Chopra A, Salomon-Escoto K, Fonseca JE, et al. How to treat patients with rheumatoid arthritis when methotrexate has failed? The use of a multiple propensity score to adjust for confounding by indication in observational studies. Ann Rheum Dis 2019;78:25-30.
18. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: From naive enthusiasm to intuitive understanding. Stat Methods Med Res 2012;21:273-93.
19. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 2015;34:3661-79.
20. Kihara M, Davies R, Kearsley-Fleet L, Watson KD, Lunt M, Symmons DPM, et al. Use and effectiveness of tocilizumab among patients with rheumatoid arthritis: An observational study from the british society for rheumatology biologics register for rheumatoid arthritis. Clin Rheumatol 2017;36:241-50.
21. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol 2012;12.
22. Almaghlouth I SJ, Pullenayegum E, Johnson S, Gladman DD, Urowitz M. Exploring the relation between immunoglobulins level and infection risk in adult patients with systemic lupus erythematosus [abstract]. Arthritis Rheumatol 2018; 70 (suppl 10).
23. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med 2013;32:2837-49.
24. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: Reporting measures of effect similar to those used in randomized experiments. Stat Med 2014;33:1242-58.
25. Elze MC, Gregson J, Baber U, Williamson E, Sartori S, Mehran R, et al. Comparison of propensity score methods and covariate adjustment: Evaluation in 4 cardiovascular studies. J Am Coll Cardiol 2017;69:345-57.
26. Gladman D, Ginzler E, Goldsmith C, Fortin P, Liang M, Urowitz M, et al. The development and initial validation of the systemic lupus international collaborating clinics/american college of rheumatology damage index for systemic lupus erythematosus. Arthritis Rheum 1996;39:363-9.
27. Urowitz MB, Ohsfeldt RL, Wielage RC, Kelton KA, Asukai Y, Ramachandran S. Organ damage in patients treated with belimumab versus standard of care: A propensity score-matched comparative analysis. Ann Rheum Dis 2019;78:372-9.
28. Urowitz MB, Su J, Gladman DD. Atherosclerotic vascular events in systemic lupus erythematosus: An evolving story. J Rheumatol 2020;47:66-71.

Table 1. Adequacy of balancing between low immunoglobulins and normal immunoglobulins groups after using PS in matching and inverse probability weighting

VARIABLE	VALUE	Normal Ig	Low Ig	STD. Diff before PS methods	STD. Diff after PS matching	STD. Diff after inverse probability weighting
		N = 656	**N = 437**		**N = 922**	**N = 1093**
Age (years)	Mean ± SD	37.69 ± 16.01	42.37 ± 14.10	0.19	0.14	0.176
Female	n (%)	388 (89)	570 (87)	0.06	0.04	0.004
Disease duration (years)	Mean ± SD	7.6 ± 8.0	11.2 ± 9.1	0.43	0.21	0.334
SLEDAI 2K	Mean ± SD	5.9 ± 5.9	6.2 ± 6.3	0.05	0.04	0.024
SDI	Mean ± SD	0.5 ± 1.0	1.2 ± 1.6	0.59	0.32	0.46
Proteinuria	n (%)	74 (11.3%)	112 (25.6%)	0.39	0.18	0.29
APA	n (%)	168 (26.2%)	62 (15.2%)	0.25	0.26	0.17
Steroid use	n (%)	349 (53.2%)	332 (76.0%)	0.48	0.14	0.31
Steroid dose (mg/day)	Mean ± SD	15.3 ± 14.6	16.8 ± 16.8	0.32	0.1	0.20
Immuno-suppressive	n (%)	152 (23.2%)	201 (46.0%)	0.5	0.17	0.36
Biologics	n (%)	1 (0.2%)	5 (1.1%)	0.13	0.01	0.09

Footnotes

STD Diff Standardized Differences, PS Propensity Score, Ig Immunoglobulin, SLEDAI Systemic Lupus Erythematosus Disease Activity Index, SDI Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index

APA Antiphospholipid antibody, SD standard deviation

Table 2. Comparison of estimates of risk of infection in SLE patients with and without low IgA immunoglobulins using 3 propensity score methods.

	Low IgA	
Propensity Score Method	**Hazard Ratio**	**95% Confidence Interval**
Matching	2.24	1.61, 3.12
Inverse Probability Weighting	1.75	1.01, 3.02
Adjustment	3.19	1.17, 8.71