Bone Marrow Transplantation (2020) 55:4–8
https://doi.org/10.1038/s41409-018-0424-x

EDITORIAL

Causal inference in randomized clinical trials

Cheng Zheng[1] · Ran Dai[2] · Robert Peter Gale[3] · Mei-Jie Zhang[4]

Received: 7 November 2018 / Accepted: 12 November 2018 / Published online: 26 March 2019

Series Editors' Note

When we do a clinical trial in which we randomize for one variable, say adding pretransplant anti-thymocyte globulin (ATG), and we see a benefit, say less graft-*versus*-host disease (G*v*HD), most people assume receiving ATG *caused* the benefit. This reasoning, termed *causal inference*, is common but wrong. Reasons why is described in the accompanying typescript. What we observe is an *association* or *correlation* between ATG and less G*v*HD, not necessarily the *cause*. This incorrect reasoning is referred to as the *association-causation fallacy*. A good example is the *correlation* between US *per capita* cheese consumption and deaths by strangulation from bedsheets with a Pearson correlation coefficient of 0.95 (see below). This and other problems of human cognition can be found in *Thinking, Fast and Slow* by Daniel Kahneman.

How can we reconcile this discordance between the goal of the clinical trialist who wants to know *why* G*v*HD is decreased and the rigor of the statistician? In the following typescript Zheng and colleagues describe the difference between *causality* and *association*. They describe statistical methods by which we can plausibly infer *causality* to results of a randomized clinical trial. We hope this typescript and others will prompt a dialogue between readers and statisticians interested in analyses of data from clinical trials of haematopoietic cell transplants. We welcome comments at #BMTStats.

Introduction

Correct interpretation of statistical data requires caution in implying *causality* [1–3], a caution contrasting with the purpose of most clinical trials whose major objective is the opposite, to assign *causality*. A typical example showing that association does not imply causation is given in Fig 1. How can we reconcile these opposing considerations? In this brief review we provide basic definitions of *causal inference* and discuss why *treatment effect*, a common clinical trial endpoint after an intervention, should not be interpreted as implying *causation*. We provide a concise guide on how to conduct statistical analyses to obtain results where *causal interpretation* may be reasonable. We also introduce classical causal methods for randomized trials and discuss methods to use covariate information to improve efficiency and methods to deal with non-compliance. Lastly, we introduce recent advanced research in causal inference of survival effects including methods for time-varying treatment experiments and high-dimensional covariate information.

Causal inference

We begin with some basic concepts in *causal inference* illustrating why it is wrong to draw conclusions regarding *causality* using *treatment effect* from simple group comparisons. Under the stable unit treatment value assumption (SUTVA) *causal effects* (or *causal estimand*[1]) are defined based on the comparison of certain functionals of the distribution of potential outcomes after two different actions

✉ Cheng Zheng
zhengc@uwm.edu

[1] Joseph. J. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

[2] Department of Statistics, University of Chicago, Chicago, IL, USA

[3] Haematology Research Centre, Division of Experimental Medicine, Department of Medicine, Imperial College London, London, UK

[4] Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA

[1] An *estimand* is a parameter which is to be estimated in a statistical analysis. The term is used to more clearly distinguish the target of inference from the function to obtain this parameter (i.e., the estimator) and the specific value obtained from a given data set.

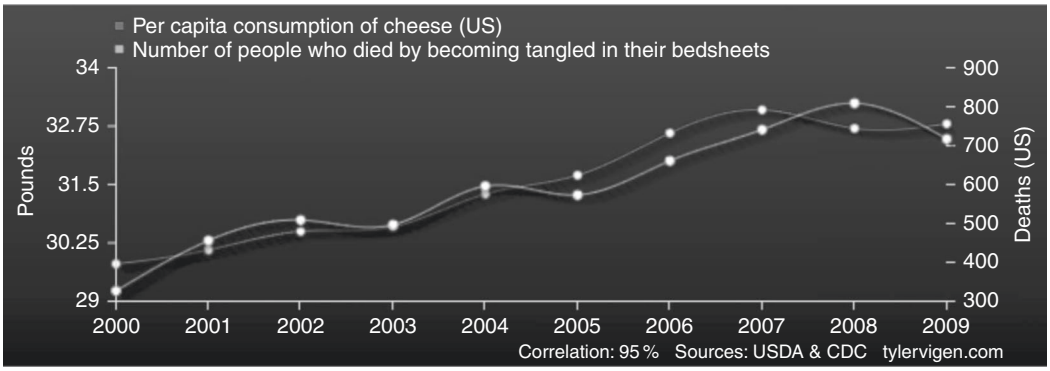

Fig. 1 Per capita consumption of cheese US. Number of people who died by becoming tangled in their bedsheets

(treatment or control) made on the same object or group of objects (for example, subjects in a clinical trial) [4, 5]. In the survival setting we are interested in quantities such as the potential survival function, cumulative hazard function (for example, cumulative incidence of relapse), restricted mean survival time (RMST) [6] and/or residual life-time [6–11]. Because treatment effect is obtained by comparing two groups it does not convey causal information. Obviously, one can never observe both potential outcomes since only one action such as treatment or placebo can be taken in each subject. This exclusivity is called *the fundamental problem of causal inference* [1, 12]. To identify causal effects considering the unavoidably missing data requires assumptions regarding the assignment mechanism of treatment and control. These different assumptions require methods other than *treatment effect* if one wants to accurately estimate causality. We expand on this point below.

Workflow for causal inference analyses

To obtain causal interpretation, we need to define the causal *estimand* through potential outcome framework (introduced in section 'Causal inference') and figure out a way to find an estimator (a functional of observed data) to identify this causal estimand. The choice of the estimator depends on the type of data we have: (1) whether there is covariate information and whether the covariate is balanced (2) whether there is non-compliance issue; and (3) whether the treatment is at one time-point or time-varying treatment. If we have one time-point treatment and the data is from a perfect randomized trial with no non-compliance, the methods from section 'Randomization methods' can be used to to make causal conclusions. However, if we also have co-variate information available and some of the co-variates are unbalanced, we can consider methods from section 'Improving efficiency with co-variate balance and adjustment to gain efficiency'. Next, we need to consider if there is non-compliance in the trial. If so, it may be necessary to use methods we discuss in section 'Non-compliance'. When there is time-varying treatment or high-dimensional covariates methods we discuss in section 'Advanced topics' should be used.

Randomization methods

Therapy-assignment conforming to individualistic, probabilistic and un-confounded assumptions is defined as a classic randomized experiment [12]. If one further assumes a constant effect we can use the Fisher exact *P*-value method under random censoring [13]. However, if we are only interested in the average causal effect (ACE) we do not need the strong constant effect assumption implied in the Fisher exact *P*-value and can use the Neyman approach by subtracting the average of treated group with the average of untreated group [12, 14]. Inverse probability of censoring weighting can be used to deal with censoring [15]. We also need to consider that in a classic randomized trial the Neyman approach provides a consistent estimator for ACE and coincides with the simple group comparison. However, this approach ignores potentially important covariate data. We discuss how we can improve on this next.

Improving efficiency with covariate balance and adjustment

Analysis of a randomized clinical trial with covariate information (for example, age, sex *etc.*) can be improved by regression adjustment and model-based imputation methods [12]. A key point in using these methods is when there are interaction terms they must be added to get an unbiased estimator for the super population ACE. If the model form is non-collapsible (for example, a Cox model), proper integration over covariate distribution is needed to compute the correct ACE [8, 16]. When sample size is small and numbers of covariates large (often so in haematopoietic cell transplant

trials), propensity score (the probability of a unit to be assigned treatment given all covariates) can be used to reduce finite sample bias and increase efficiency [17]. Several propensity score-based methods are available including propensity score matching for sub-classification [18–22], propensity score adjustment [23], trimming based on propensity score [24] and variants combining several techniques [15]. In practice, the propensity score is unknown and is commonly fitted from a logistic regression model. To deal with issues from potential model misspecification, the multiple robust estimator [15] can be used to analyze the data at the price of potential loss of efficiency.

Non-compliance

Non-compliance is common in clinical trials and makes implying *causal inference* even more difficult. A common practice is to analyze treatment effect by *intent-to-treat* ignoring compliance such that randomization assumptions still operate. In this way methods in sections 'Workflow for causal inference analyses' and 'Randomization methods' are all valid to estimate the ACE of the *intent-to-treat* effect. This approach is obviously different from the ACE of the real treatment because it includes subjects not receiving the assigned treatment but analyzed as if they had. A less valid approach to estimating treatment effect is to analyze only data from subjects assigned to and receiving the therapy. However, this approach violates the randomization assumptions. Consequently, methods we describe in sections 'Workflow for causal inference analyses' and 'Randomization methods' cannot lead to an unbiased estimator for ACE.

A proposed solution to this problem is the principal stratification method [25]. Subjects are classified into four latent groups based on their potential compliance state under different treatment assignments: (1) complier; (2) always taker; (3) never taker; and (4) defier. A subject's group attribute can only be partially identified directly from his observed compliance state (for example, a subject who complied to the treatment is either a complier or an always taker). Treatment efficacy is usually considered as the ACE in the complier group, identifiable based on different combination of assumptions such as exclusion criteria, monotonicity and/or parametric model assumption [26–28]. Or we can obtain a bound for the *causal effect* with weaker assumptions [29–31]. Another way to handle non-compliance is to consider compliance state as a mediator and use an instrumental variable approach to handle potential un-measured confounders between compliance and outcome [32–36]. Another way is to assume sequential ignorability. Under this assumption the true treatment action (whether the subject received the therapy or not) can be analyzed as conditionally randomized and therefore propensity score methods or methods specifically designed for non-compliance issue [37–40] can be used to estimate causality followed by a sensitivity analyses [41–45] to evaluate robustness of the estimator under assumption violation.

Advanced topics

There is considerable recent research in how to estimate *causal inference* in survival data analyses. One important direction is how to deal with the time-varying treatment studies. These are studies where the treatment is not a one-time binary choice (for example, a transplant versus chemotherapy) [46] but assigned over time and possibly adjusted based on prior outcomes (for example, giving azacitidine to subjects with a positive posttransplant measurable residual disease [MRD]-test). Cox models using time-dependent covariates face the problem of confounding such as prior therapy treatment and cannot therefore imply causality. The marginal structural model [47–49] and the nested structural mean model [50, 51] should be considered for these types of data. Another important direction is how to take advantage of rapidly-increasing availability of covariate data from clinical trials to increase efficiency of ACE estimations. High-dimensional methods using novel machine learning techniques have been developed for this purpose [52–54].

Conclusion

In the brief review, we provided the concept of the potential outcome framework and described the complex challenge of inferring causality in survival analyses of randomized clinical trials. We discuss limitations of estimating causality under these conditions and suggest potential statistical techniques to help estimate causality with greater accuracy.

Acknowledgements RPG acknowledges support from the National Institute of Health Research (NIHR) Biomedical Research Centre funding scheme. MJZ acknowledges support from the National Institute of Health (NCI, NHIBI) and Health Resources and Services Administration (HRSA).

Compliance with ethical standards

Conflict of interest RPG is a part-time employee of Celgene Corp. The remaining authors declare that they have no conflict of interest.

References

1. Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986;81:945–60.
2. Aldrich J. Correlations genuine and spurious in Pearson and Yule. (1995). Stat Sci. 1995;10:364–76.
3. Pearl J. Causality: models, reasoning, and inference. 2nd edn. Cambridge: Cambridge University Press; 2009.
4. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.
5. Rubin D. Causal inference using potential outcomes. J Am Stat Assoc. 2005;100:322–31.
6. Chen P, Tsiatis AA. Causal inference on the difference of the restricted mean life between two groups. Biometrics. 2001;57:1030–8.
7. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol. 2014;32:2380–5.
8. Hernan MA. The hazards of hazard ratios. Epidemiology. 2010;21:13–15.
9. Lin H, Li Y, Jiang L, Li G. A semiparametric linear transformation model to estimate causal effects for survival data. Can J Stat. 2014;42:18–35.
10. Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Anal. 2015;21:579–93.
11. Vock DM, Tsiatis AA, Davidian M, Laber EB, Tsuang WM, Finlen-Copland A, et al. Assessing the causal effect of organ transplantation on the distribution of residual lifetime. Biometrics. 2013;69:820–9.
12. Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences. Cambridge: Cambridge University Press; 2015.
13. Rosenbaum P. Conditional permutation tests and the propensity score in observational studies. J Am Stat Assoc. 1984;79:565–74.
14. Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc. 1934;97:558–625.
15. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc. 1995;90:106–21.
16. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Anal. 2013;19:279–96.
17. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
18. Rubin DB, Thomas N. Matching using estimated propensity score: relating theory to practice. Biometrics. 1996;52:249–64.
19. Rubin DB, Thomas N. Combining propensity score matching with additional adjustment for prognostic covariates. J Am Stat Assoc. 2000;95:573–85.
20. 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.
21. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968;24:295–313.
22. Gu X, Rosenbaum P. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat. 1993;2:405–20.
23. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Statistics in Medicine. 2013;32: 2837–49.
24. Crump R, Hotz VJ, Imbens G, Mitnik O. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96:187–99.
25. Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58:21–29.
26. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. Ann Stat. 1997;25:305–27.
27. Cuzick J, Sasieni P, Myles J, Tyrer J. Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. J R Stat Soc, Ser B. 2007;69:565–88.
28. Little RJ, Long Q, Lin X. A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. Biometrics. 2009;65:640–9.
29. Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. J Am Stat Assoc. 1997;92:1171–6.
30. Cheng J, Small DS. Bounds on causal effects in three-arm trials with non-compliance. J R Stat Soc, Ser B. 2006;68:815–36.
31. Heckman JJ, Vytlacil EJ. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc Natl Acad Sci USA. 1999;96:4730–4.
32. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). J Am Stat Assoc. 1996;91:444–72.
33. Tchetgen Tchetgen EJ, Walter S, Vansteelandt S, Martinussen T, Glymour M. Instrumental variable estimation in a survival context. Epidemiology. 2015;26:402–10.
34. Zheng C, Dai R, Hari PN, Zhang MJ. Instrumental variable with competing risk model. Stat Med. 2017;36:1240–55.
35. Li J, Fine J, Brookhart A. Instrumental variable additive hazards models. Biometrics. 2015;71:122–30.
36. Martinussen T, Nøbro Sørensen D, Vansteelandt S. Instrumental variables estimation under a structural Cox model. Biostatistics. 2019;20:65-79.
37. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56:779–88.
38. Loeys T, Goetghebeur E. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. Biometrics. 2003;59:100–5.
39. Frangakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika. 1999;86:365–79.
40. Robins JM, Tsiatis AA. Correcting for noncompliance in randomized trials using rank preserving structural failure time models. Commun Stat. 1991;20:2609–31.
41. VanderWeele TJ. Unmeasured confounding and hazard scales: sensitivity analysis for total, direct, and indirect effects. Eur J Epidemiol. 2013;28:113–7.
42. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics. 1998;54:948–63.
43. VanderWeele TJ. Sensitivity analysis: distributional assumptions and confounding assumptions. Biometrics. 2008;64:645–9.
44. Carnegie NB. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. J Res Educ Eff. 2016;9:395–420.
45. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. Epidemiology. 2016;27:368–77.
46. Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. Stat Med. 2002;21:2449–63.

47. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000;11:561–70.
48. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–60.
49. Bodnar L, Davidian M, Siega-Riz AM, Tsiatis AA. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. Am J Epidemiol. 2004;159:926–34.
50. Robins JM. Structural nested failure time models. In: Armitage P, Colton T (eds) Encyclopedia of Biostatistics (Chichester: John Wiley & Sons., 1998) pp. 4372–89.
51. Vansteelandt S, Joffe M. Structural nested models and G-estimation: the partially realized promise. Stat Sci. 2014;29: 707–31.
52. van der Laan MJ, Rose S. Targeted learning causal inference for observational and experimental data. New York: Springer; 2011.
53. Wager S, Du W, Taylor J, Tibshirani RJ. High-dimensional regression adjustments in randomized experiments. Proc Natl Acad Sci USA. 2016;113:12673–8.
54. Bloniarz A, Liu H, Zhang CH, Sekhon JS, Yu B. Lasso adjustments of treatment effect estimates in randomized experiments. Proc Natl Acad Sci USA. 2016;113:7383–90.