Stephen R. Cole*, Michael G. Hudgens and Jessie K. Edwards

A Fundamental Equivalence between Randomized Experiments and Observational Studies

DOI 10.1515/em-2015-0029

Abstract: A fundamental probabilistic equivalence between randomized experiments and observational studies is presented. Given a detailed scenario, the reader is asked to consider which of two possible study designs provides more information regarding the expected difference in an outcome due to a time-fixed treatment. A general solution is described, and a particular worked example is also provided. A mathematical proof is given in the appendix. The demonstrated equivalence helps to clarify common ground between randomized experiments and observational studies, and to provide a foundation for considering both the design and interpretation of studies.

Keywords: randomized trial, observational study, bounds

Below we describe a fundamental probabilistic equivalence between randomized experiments and observational studies. This equivalence clarifies the logical common ground between randomized experiments and observational studies. This equivalence is one of equal bounds. Bounds are the minimum and maximum values of the parameter (e. g., risk difference) that are consistent with the observed data distribution. Such bounds were described by Robins (1989), Manski (1990), and Balke and Pearl (1994). Moreover, this equivalence illustrates a fundamental limitation of observational studies, as compared to randomized experiments.

1 Two study designs

Say you are charged with the task of learning if and to what extent a binary, time-fixed treatment, denoted as A with levels $a = \{0, 1\}$, affects a binary outcome of interest, measured at some fixed time-point after treatment and denoted as Y with values y. Your goal is to produce an estimate of the expected difference in the outcome Y due to the treatment (e. g., risk difference (Cole et al. 2015). With limited resources, you must choose one of the two following study designs. You will want to choose the design that provides more information regarding this expected difference in the outcome Y due to the treatment. By "information" here we mean the width of the bounds for the risk difference.

The first study design is a randomized experiment (Fisher 1926). Specifically, you randomly assign n participants with proportion p to treatment $A = 1$ and proportion $1 - p$ to treatment $A = 0$. Say a complication is that under this first study design exactly half of the outcome assessments will be missing, balanced equally across treatment groups. To emphasize, in this thought experiment, for illustrative purposes, half the outcome data are missing for the randomized experiment. The second study design is an observational (cohort) study. You observe n participants where (the same) proportion p choose treatment $A = 1$ and proportion $1 - p$ choose treatment $A = 0$. Under this study design you observe all n outcome assessments.

Assume n is so large, and that the participants are a random sample from the target population, such that we can ignore the issues of random error and generalizability, respectively. Assume the outcome is

***Corresponding author: Stephen R. Cole,** Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA, E-mail: cole@unc.edu
Michael G. Hudgens, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA
Jessie K. Edwards, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA

measured without error in both studies. Assume that in the experiment there is complete compliance with assigned treatment, and in the observational study treatment is measured without error. Assume that there is no interference, i. e., one participant's treatment does not affect the outcomes of another participant. Assume there are not different versions of treatment, or treatment variation irrelevance (Cole and Frangakis 2009; Vander Weele 2009). Finally, assume that each study is the same in terms of cost and other operational issues. Which study design would provide more information (i. e., tighter bounds)?

2 Fundamental equivalence

Without additional context, both study designs presented provide the exact same amount of information (i. e., width of bounds) about the average difference in Y due to treatment! To see this point clearly, define hypothetical values of Y under each treatment, which are called potential outcomes (Holland 1986; Neyman et al. 1990; Robins 1986; Rubin 1974). The difference in Y due to treatment is the difference between the value of Y if a participant had been assigned (or chosen) treatment $A = 1$, denoted Y^1, and the value of Y if a participant had been assigned (or chosen) treatment $A = 0$, denoted Y^0. Define the expected difference in the outcome Y in the target population due to the treatment by $E(Y^1 - Y^0)$. A fundamental problem of causal inference is that we never observe both Y^1 and Y^0 for the same participant. If we did observe Y^1 and Y^0 for each participant, then the average of $Y^1 - Y^0$ would be an unbiased estimate of $E(Y^1 - Y^0)$.

For both studies, when we observe Y for a participant with $A = a$, then $Y^a = Y$ by counterfactual consistency (Pearl 2010), under our assumptions of no exposure measurement error and no versions of treatments, or treatment-version irrelevance (Cole and Frangakis 2009; Vander Weele 2009). Imagine for a moment that the first (experimental) study design was implemented with no missing data. Then counterfactual consistency allows us to observe 1/2 of the potential outcomes, namely Y^a for $A = a$. The average of these observed outcomes for individuals assigned $A = 1$ is an unbiased estimator of $E(Y^1|A = 1)$, which equals $E(Y^1)$ because treatment assignment is independent of the potential outcomes by randomization. Similarly the average of the observed outcomes for individuals assigned $A = 0$ is an unbiased estimator of $E(Y^0|A = 0)$, which by randomization equals $E(Y^0)$. Thus the difference in average outcomes between the two arms of the study is an unbiased estimator of the treatment effect $E(Y^1) - E(Y^0) = E(Y^1 - Y^0)$. This is a central reason randomized designs are so highly regarded. Unfortunately, however, for the randomized experiment under consideration, half of the outcomes assessments are missing. The mechanism by which outcome assessments are missing is unknown.

In the second (observational) study design, counterfactual consistency again allows us to observe 1/2 of the potential outcomes, namely Y^a for $A = a$. But we do not observe the other half of the potential outcomes, namely Y^a for $A \neq a$. Because the treatment was not randomized, the mechanism by which individuals chose treatment is unknown.

Because potential outcomes are missing due to an unknown mechanism under either design we cannot with certainty obtain an unbiased estimator of the expected difference in the outcome Y due to treatment, $E(Y^1 - Y^0)$. However, with the observed data, we can calculate bounds on $E(Y^1 - Y^0)$ under each design. These bounds are the same for each of the two study designs, proving the studies provide the same information about the effect of treatment. A mathematical proof of the equivalence of the bounds is provided in the Appendix.

3 An example

Table 1 provides a simple worked numerical example. In the upper panel of Table 1 the experimental study is represented. There are n participants, with $0.5n$ assigned to $A = 1$ and $0.5n$ assigned to $A = 0$. Among the

Table 1: Illustrative example of a randomized experiment and observational study.

Randomized experiment:

	Missing outcomes $M=1$	Observed nonevents $M=0, Y^a=0$	Observed events $M=0, Y^a=1$	Total
Treatment				
$A=1$	0.25	0.225	0.025	0.5
$A=0$	0.25	0.200	0.050	0.5
Total	0.50	0.425	0.075	1.0

Observed risk difference is: 0.025/0.25 − 0.05/0.25 = 0.1 − 0.2 = − 0.1
Bounds are: −0.55, 0.45

Observational study:

	Nonevents $Y=0$		Events $Y=1$		Total
	$Y^0=0$	$Y^1=0$	$Y^0=1$	$Y^1=1$	
Treatment					
$A=1$	?	0.45	?	0.05	0.5
$A=0$	0.40	?	0.10	?	0.5
Total					1.0

Observed risk difference is: 0.05/0.5 − 0.1/0.5 = 0.1–0.2 = −0.1
Bounds are: −0.55, 0.45.

$0.5n$ with observed outcome status, there are $0.025n$ treated events (i. e., $A=1, Y=1$) and $0.05n$ untreated events (i. e., $A=0, Y=1$), yielding a risk difference of − 0.1. Accounting for the $0.5n$ participants with missing outcomes, the logically possible range (i. e., bounds) of the risk difference are − 0.55 and 0.45. Specifically, the lower bound occurs if we assume all participants with missing outcomes assigned to $A=1$ are non-events (i. e., $Y=0$) and all participants with missing outcomes assigned to $A=0$ are events, or $0.025/0.5 - 0.3/0.5 = 0.05 - 0.6 = -0.55$. The upper bound occurs if we assume all treated participants with missing outcomes are events and untreated participants with missing outcomes are non-events, or $0.275/0.5 - 0.05/0.5 = 0.55 - 0.1 = 0.45$.

In the lower panel of Table 1 the observational (cohort) study is represented. Again, there are n participants, with $0.5n$ treated $(A=1)$ and $0.5n$ untreated $(A=0)$. Here every factual outcome Y is observed but the counterfactual outcomes Y^a for $A \neq a$ are missing. Accounting for the n missing potential outcomes, the bounds of the risk difference are again − 0.55 and 0.45. The lower bound occurs if we assume all untreated participants would not have experienced an event if, contrary to fact, they had been treated, and all treated participants would have experienced the event if, contrary to fact, they been untreated. In this scenario, the $0.05n$ events observed among participants with $A=1$ are the only events that would have been observed had all participants been treated (i. e., $P(Y^1=1)=0.05$) and the $0.1n$ events observed among participants with $A=0$ would be added to the $0.5n$ events we would have observed among participants with $A=1$, had they been untreated (i. e., $P(Y^0=1)=0.6$), for a risk difference of $0.05 - 0.6 = -0.55$.

The upper bound occurs if we assume all untreated participants would have experienced an event if, contrary to fact, they had been treated, and all treated participants would have not experienced the event if, contrary to fact, they had been untreated. Under this scenario, we assume that all $0.5n$ untreated participants would have had the event if (contrary to fact) they had been treated in addition to the $0.05n$ events we observed among participants with $A=1$ participants, so $P(Y^1=1)=0.55$. Similarly, we observed $0.1n$ events among participants with $A=0$ and assume no treated participants would have had events if (contrary to fact) they had been untreated, so $P(Y^0=1)=0.1$, making the upper bound of the risk difference $0.55 - 0.1 = 0.45$.

4 Discussion

With a pair of treatments and assuming no additional context beyond what is provided above, the bounds for the risk difference from an observational study are equivalent to the analogous bounds for a randomized experiment with 50 % missing outcomes. Of course the described experiment may be labeled as "broken" because of the missing outcomes (Frangakis and Rubin 1999; Little et al. 2012). However, experience suggests that all real-world experiments are broken to varying degrees. In real-world settings each successive additional piece of context will unbalance the equivalence given in this example, giving preference to one design over the other. A central point of this paper is that we cannot conclude one design is better than the other without additional context. Such additional context, could be used to "unbalance" the equivalence and allow for an informed design choice. Yet identifying this balancing point, or equivalence, between a randomized experiment with missing outcome assessments and an observational study is useful. When a randomized experiment is feasible, this equivalence indicates that an experiment will be preferable to an observational study provided less than 50 % of outcome assessments in the experiment are missing. However, experiments are sometimes unethical and are often prohibitively expensive. When a randomized experiment is infeasible, nonexperimental observational studies can help to refine our knowledge, or sharpen our (probabilistic) bounds about the effect of a treatment. Moreover, balancing points, such as described here, provide a natural foundation when considering both the design and interpretation of experimental and nonexperimental studies.

Funding: Dr. Cole was supported in part by grants R01AI100654, R24AI067039, U01AI103390, and P30AI50410 from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix: Proof

For each design we will calculate the two bounds for $E(Y^0)$. The same approach can be used for $E(Y^1)$. We may then subtract the lower bound for $E(Y^0)$ from the upper bound for $E(Y^1)$ to calculate the upper bound for the risk difference, $E(Y^1 - Y^0)$. The lower bound for $E(Y^1 - Y^0)$ can be computed analogously.

First consider the randomized experiment where

$$E\left(Y^0\right) = E\left(Y^0|A=0\right) = E(Y|A=0).$$

In the above, the first equality holds by randomization of treatment A and the second equality holds by counterfactual consistency. The far right term is identified from the observed data when there are no missing outcome assessments. However, for the experiment described in the main text half of the outcome assessments were missing. Let $M=1$ indicate the outcome assessment is missing, $M=0$ otherwise. We may expand $E(Y|A=0)$ as

$$E(Y|A=0,M=0)P(M=0|A=0) + E(Y|A=0,M=1)P(M=1|A=0).$$

For the experiment described in the main text, missingness does not depend on treatment group, implying $P(M=m|A=0) = P(M=m)$for $m=0,1$. Therefore we can simplify the above as

$$E(Y|A=0,M=0)P(M=0) + E(Y|A=0,M=1)P(M=1).$$

The first, second and fourth terms are identified in the observed data, but the third term is not. Therefore, setting the unidentified third term $E(Y|A=0,M=1)$ to equal the minimum possible value of 0, yields the lower bound on $E(Y|A=0)$, namely $E(Y|A=0,M=0)P(M=0)$. Alternatively, setting the unidentified third term $E(Y|A=0,M=1)$ equal to the maximum possible value of 1, yields the upper bound on $E(Y|A=0)$, namely $E(Y|A=0,M=0)P(M=0)+P(M=1)$. The difference between the lower and upper bounds is

$P(M=1)$. Likewise, the difference between the lower and upper bounds for $E(Y^1)$ is $P(M=0)$. Therefore, whatever the value of $P(M=1)$ the length of the bounds on the risk difference is 1 because, by convexity, $P(M=1)+P(M=0)=1$.

Likewise, in the observational study we can expand $E(Y^0)$ as

$$E\left(Y^0|A=0\right)P(A=0)+E\left(Y^0|A=1\right)P(A=1).$$

By counterfactual consistency we have

$$E(Y|A=0)P(A=0)+E\left(Y^0|A=1\right)P(A=1).$$

Again, the first, second and fourth terms are identified in the observed data, but the third term is not. Therefore, setting the unidentified third term $E(Y^0|A=1)$ equal to the minimum possible value of 0, yields the lower bound on $E(Y^0)$, namely $E(Y|A=0)P(A=0)$. Alternatively, setting the unidentified third term $E(Y^0|A=0)$ equal to the maximum possible value of 1, yields the upper bound on $E(Y^0)$, namely $E(Y|A=0)P(A=0)+P(A=1)$. The difference between the lower and upper bounds is $P(A=1)$. Likewise, the difference between the lower and upper bounds for $E(Y^1)$ is $P(A=0)$. Therefore, whatever the value of $P(A=1)$ the length of the bounds on the risk difference is 1 because, by convexity, $P(A=1)+P(A=0)=1$.

References

Robins, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Health Service Research Methodology: A Focus on AIDS, L. Sechrest, H. Freeman, and A. Mulley (Eds.), 113–159. Washington, DC: US Public Health Service.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. The American Economic Review 80:319–323.

Balke, A., and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds, and applications. In: Uncertainty in Artifical Intelligence, R. Lopez de Mantara, and D. Poole (Eds.), 46–54. San Mateo, CA: Morgan Kaufman.

Cole, S. R., Hudgens, M. G., Brookhart, M. A., Westreich, D. (2015). Risk. American Journal of Epidemiology, 181(4):246–250.

Fisher, R. A. (1926). The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain, 33:503–513.

Cole, S. R., and Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? Epidemiology, 20(1):3–5.

Vander Weele, T. J. (2009). Concerning the consistency assumption in causal inference. Epidemiology, 20(6):880–883.

Neyman, J., Dabrowska, D. M., Speed, T. P. (1990). On the application of probability theory to agricultural experiments: Essay on principles, section 9 (1923). Statistical Science, 5:465–480.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66:688–701.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period: Application to control of the healthy worker survivor effect. Math Modelling, 7:1393–1512.

Holland, P. W. (1986). Statistics and causal inference. JASA, 81:945–970.

Pearl, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? Epidemiology, 21(6):872–875.

Frangakis, C. E., and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika, 86:365–379.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P., Stern, H. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14):1355–1360.