Practice of Epidemiology

Using Bounds to Compare the Strength of Exchangeability Assumptions for Internal and External Validity

Alexander Breskin*, Daniel Westreich, Stephen R. Cole, and Jessie K. Edwards

* Correspondence to Dr. Alexander Breskin, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, McGavran-Greenberg Hall, CB 7435, Chapel Hill, NC 27599 (e-mail: abreskin@live.unc.edu).

Initially submitted August 21, 2018; accepted for publication February 27, 2019.

In the absence of strong assumptions (e.g., exchangeability), only bounds for causal effects can be identified. Here we describe bounds for the risk difference for an effect of a binary exposure on a binary outcome in 4 common study settings: observational studies and randomized studies, each with and without simple random selection from the target population. Through these scenarios, we introduce randomizations for selection and treatment, and the widths of the bounds are narrowed from 2 (the width of the range of the risk difference) to 0 (point identification). We then assess the strength of the assumptions of exchangeability for internal and external validity by comparing their contributions to the widths of the bounds in the setting of an observational study without random selection from the target population. We find that when less than two-thirds of the target population is selected into the study, the assumption of exchangeability for external validity of the risk difference is stronger than that for internal validity. The relative strength of these assumptions should be considered when designing, analyzing, and interpreting observational studies and will aid in determining the best methods for estimating the causal effects of interest.

causal inference; external validity; internal validity; partial identification

The goal of many epidemiologic studies is to estimate the causal effect of an exposure or treatment (henceforth called treatment) on an outcome in a population of interest (the target population) (1). One widely accepted definition of a causal effect involves a comparison of the distributions of potential outcomes for different treatment scenarios (2–4). Each potential outcome corresponds to the outcome a subject would experience if, possibly counter to fact, he or she received a given level of the treatment. However, because each subject has only one treatment experience, in studies comparing mutually exclusive treatment plans, at most half of these potential outcomes can be observed (an issue often referred to as the fundamental problem of causal inference (5)), and thus the point identification of causal effects often relies on the assumption that a set of sufficient identification conditions are met.

One condition that is perhaps most familiar to epidemiologists is exchangeability (specifically, exchangeability between participants receiving each level of treatment). This is the condition that the potential outcomes are independent of observed treatment, possibly conditional on a set of covariates (6). While methods to aid in identifying a sufficient set of covariates have been developed (7), it is impossible to verify that exchangeability between persons receiving different levels of treatment holds.

Without assuming exchangeability between treatment groups, then, even assuming perfect measurement of exposure, outcome, and covariates, the causal effect is only set-identified, which means that we can only produce bounds, rather than a single-point solution (8–10).

In most settings, complete data on the target population are unavailable. In that case, all of the potential outcomes are unobservable for persons who are not included in the study population. While lack of exchangeability between treatment groups is a concern for *internal validity*, lack of exchangeability between the study and target populations is a concern for *external validity* (11). In recent work, Westreich et al. (1) highlighted the role that external validity plays in causal effect estimation and suggested that internal and external validity be considered together with the concept of *target validity*. Without assuming exchangeability between the study population and the target population, the causal effect is again only set-identified, so only bounds on the effect of interest can be produced—an issue referred to by Manski (12) as the selection problem.

With no data and no assumptions, all that can be said is that the risk difference is bounded by its mathematical range, $[-1, 1]$. With data or assumptions, these bounds can be narrowed substantially, and descriptions of such bounds for many cases have been

described previously (8–10, 12–14). In this paper, we describe bounds that do not require an exchangeability assumption for 4 specific scenarios that researchers may encounter: observational studies and randomized studies, each with and without simple random selection from the target population. Through these scenarios, we introduce randomizations of selection and treatment, eventually reducing the width of the bounds from 2 (the width of the range of the risk difference) to 0 (point identification). We then assess the relative strength of the exchangeability assumptions for internal and external validity by comparing the widths of the bounds attributable to each in the setting of an observational study without random selection from the target population. Throughout this paper, we do not consider the random variability of the bounds, though we direct the reader to methods for incorporating such random variability in the Discussion section.

NOTATION

Persons in the target population will be indexed by i , $i \in \{1, 2, \dots, n\}$. The treatment and outcome of interest are binary and are defined as $A_i = 1$ if subject i is treated (0 otherwise) and $Y_i = 1$ if subject i experiences the outcome (0 otherwise). We only have information on a subset of the target population, and we define $S_i = 1$ if subject i is included in the study population (0 otherwise). The study population may be a subset of the target population (in the case of generalizing from the study population to the target) or not (in the case of transporting from the study population to the target). Potential outcomes are denoted by superscripts (e.g., $Y_i^a = 1$ if subject i would experience the outcome if he or she received treatment level a).

PARAMETER OF INTEREST AND CAUSAL IDENTIFIABILITY CRITERIA

The focus of this paper is the effect of a binary treatment on a binary outcome. The parameter of interest is the causal risk difference in the target population, defined as

$$E[Y^1] - E[Y^0],$$

which has a range of $[-1, 1]$. Without data, the information we have about the risk difference is that it falls within this range, which has a width of 2.

As described by Lesko et al. (15), one sufficient set of criteria for identifying the causal risk difference from observational data with a census of the target population includes 1) exchangeability, possibly conditional on covariates, between persons receiving different levels of treatment; 2) a nonzero probability of treatment within each stratum of the covariates; 3) causal consistency; and 4) no measurement error. When not all subjects from the target population are included in the study population, additional criteria are needed, including 5) exchangeability, possibly conditional on covariates, between the members of the study and target populations; 6) a nonzero probability of being selected into the study population within each level of the covariates; 7) similar versions of treatment in the study and target populations; and 8) similar interference patterns in the study and target populations.

In this paper, we specifically consider conditions 1 and 5, and throughout this article we assume the other conditions to hold. We also assume that there are no missing data in the study population.

To be precise in our discussion, we define marginal exchangeability between persons receiving different levels of treatment within the study population as

$$Y^a \perp\!\!\!\perp A \mid S = 1; a \in \{0, 1\},$$

which means that the potential outcomes (Y^1 and Y^0) are independent of treatment (A) among those in the study population ($S = 1$). We refer to this condition as exchangeability for internal validity.

Similarly, we define marginal exchangeability between the members of the study and target populations as

$$Y^a \perp\!\!\!\perp S; a \in \{0, 1\}.$$

This means that the potential outcomes are independent of membership in the study population. We refer to this condition as exchangeability for external validity. We note that the exchangeability conditions presented here can be weakened to only hold within levels of covariates, but the results that follow remain the same.

BOUNDS FOR THE CAUSAL RISK DIFFERENCE UNDER 4 DIFFERENT SCENARIOS

An observational study without random selection from the target population

For an observational study conducted in a study population which was not randomly selected from the target population, exchangeability for both internal and external validity, which together comprise target validity (1), are concerns. If these exchangeability conditions do not hold, the resulting estimators may suffer from both internal and external validity bias (also known as target validity bias (1)).

Unfortunately, it is impossible to verify that these exchangeability assumptions hold. Without assuming they hold, only bounds on the causal effect in the target population can be identified (9, 10, 12). In this case, the bounds for the causal effect in the target population under no exchangeability assumptions are

$$\left[\begin{array}{l} E[Y|A = 1, S = 1]P(A = 1|S = 1)P(S = 1) - \\ E[Y|A = 0, S = 1]P(A = 0|S = 1)P(S = 1) - \\ P(A = 1|S = 1)P(S = 1) - P(S = 0) \end{array} \right],$$

$$\left[\begin{array}{l} E[Y|A = 1, S = 1]P(A = 1|S = 1)P(S = 1) - \\ E[Y|A = 0, S = 1]P(A = 0|S = 1)P(S = 1) + \\ P(A = 0|S = 1)P(S = 1) + P(S = 0) \end{array} \right]$$

(see derivation in the Web Appendix, available at <https://academic.oup.com/aje>). These bounds have width $2 \times P(S = 0) + P(S = 1) = 2 - P(S = 1)$, which can be seen by subtracting the lower bound from the upper bound. Note that as $P(S = 1)$ approaches 0 (meaning no members of the target

population are observed), the width of the bounds approaches 2 (the full width of the range of the risk difference).

A randomized trial without random selection from the target population

In the case of a randomized trial conducted in a study population which was not randomly selected from the target population, randomization of treatment will, in expectation, provide exchangeability between the treated and untreated subjects in the study population. Exchangeability for external validity is thus the primary concern. If exchangeability for external validity does not hold, the resulting estimators may suffer from external validity bias (1).

Again, only bounds for the causal effect in the target population are identifiable without assuming exchangeability for external validity. Following the derivation provided by Manski (12) for the selection problem, these bounds are

$$\left[\begin{array}{l} E[Y|A = 1, S = 1] - E[Y|A = 0, S = 1] - P(S = 0), \\ E[Y|A = 1, S = 1] - E[Y|A = 0, S = 1] + P(S = 0) \end{array} \right]$$

(derivation in Web Appendix). Here, the width of the bounds is $2 \times (1 - P(S = 1))$. Again, if no members of the target population are observed, the bounds span the full range of the risk difference. Conversely, if the target population is defined as the study population, then $P(S = 1) = 1$, and thus the width of the bounds is 0.

An observational study with random selection from the target population

When an observational study is conducted in a study population randomly selected from the target population, exchangeability for internal validity is the primary concern, since random selection will, in expectation, ensure exchangeability between the members of the study and target populations. If exchangeability for internal validity does not hold, the resulting estimators may suffer from internal validity bias (1).

Adapting the derivations provided by Manski (9) and Cole et al. (10), the bounds for the causal effect in an observational study randomly selected from the target population are

$$\left[\begin{array}{l} E[Y|A = 1, S = 1]P(A = 1|S = 1) - E[Y|A = 0, S = 1] \\ \times P(A = 0|S = 1) - P(A = 1|S = 1), \\ E[Y|A = 1, S = 1]P(A = 1|S = 1) - E[Y|A = 0, S = 1] \\ \times P(A = 0|S = 1) + P(A = 0|S = 1) \end{array} \right]$$

(derivation in Web Appendix). The width of the bounds in this scenario is always $P(A = 0|S = 1) + P(A = 1|S = 1) = 1$. This is coherent with the previous work by Cole et al. (10) and clarifies and highlights the fact that previous derivations of the bounds for an observational study were (as the authors noted) predicated on an assumption of random selection from the target population.

A randomized trial with random selection from the target population

If treatment is randomized and the study population is randomly selected from the target population, then in expectation random selection will provide exchangeability between the members of the study and target populations, and randomization of treatment will ensure exchangeability between the treated and untreated subjects. In this case, it is possible to point-identify the causal effect of interest as simply the difference in the risk of the outcome between the treated and untreated subjects in the study population, and the bounds have width 0 (derivation in Web Appendix).

COMPARING THE STRENGTH OF EXCHANGEABILITY ASSUMPTIONS FOR INTERNAL AND EXTERNAL VALIDITY

The width of the bounds for the causal effect estimated from an observational study with participants not randomly selected from the target population has contributions from uncertainty regarding exchangeability for both internal and external validity. The width of the bounds attributable to not assuming that each condition holds is a useful measure of the strength of the assumptions. Intuitively, the width of the bounds is proportional to the number of potential outcomes identified by leveraging the assumptions. Comparing the widths of the bounds attributable to each thus provides insight as to how much information each assumption carries.

For a binary treatment, with no assumptions, the bounds span the range of the risk difference, $[-1, 1]$, and have a width of 2. The conditions comprising causal consistency (16–18) identify one potential outcome for each member of the study population, so for a target population of n individuals, $(2 - P(S = 1))n$ potential outcomes remain unidentified, and the total width of the bounds is

$$\frac{(2 - P(S = 1))n}{2n} \times 2 = 2 - P(S = 1)$$

(note that in the case of an exposure with k levels, there are kn total potential outcomes, and $(k - P(S = 1))n$ remain unidentified after applying causal consistency). The assumption of exchangeability for internal validity allows us to identify the average of the potential outcomes under no treatment for the treated group and under treatment for the untreated group. This means that the assumption identifies the average of $P(S = 1) \times n$ potential outcomes. The assumption of exchangeability for external validity identifies the averages of *both* potential outcomes for the unselected subjects, for a total of $2 \times (1 - P(S = 1)) \times n$ potential outcomes ($k \times [1 - P(S = 1)] \times n$ for a k -level exposure). Because the width of the bounds attributable to each condition is proportional to the fraction of the potential outcomes identified by leveraging that assumption, the width attributable to not assuming exchangeability for internal validity is

$$\frac{P(S = 1) \times n}{2n} \times 2 = P(S = 1),$$

Figure 1. Contribution of exchangeability for internal and external validity to the width of bounds for the causal effect of a binary treatment on a binary outcome. Solid line, total width of the bounds; dashed line, contribution due to not assuming exchangeability for external validity; dotted line, contribution due to not assuming exchangeability for internal validity.

and the width attributable to not assuming exchangeability for external validity is

$$\frac{2 \times (1 - P(S = 1)) \times n}{2n} \times 2 = 2 \times (1 - P(S = 1)).$$

The assumptions of exchangeability for internal and external validity thus carry equal information when two-thirds of the target population is selected into the study. When less than two-thirds of the target population is selected into the study, the assumptions for external validity are stronger, and vice versa. Figure 1 shows the relative contributions of each assumption as a function of the proportion of the target population selected.

Note that the width of the bounds due to not assuming exchangeability for internal validity described in this section, $P(S = 1)$, is potentially narrower than that under the scenario of an observational study in which the study population is randomly selected from the target population, in which case the bounds have width 1, all attributable to not assuming exchangeability for

internal validity. This can be understood as follows. First, causal consistency identifies the potential outcomes for each selected subject's observed level of treatment, for a total of $P(S = 1) \times n$ potential outcomes. Next, random selection means that these distributions are the same in the unselected population, so the averages of half, or $(1 - P(S = 1)) \times n$, of the potential outcomes in the unsampled population are identified. Finally, exchangeability for internal validity identifies the averages for the remaining $2n - \{P(S = 1) + [1 - P(S = 1)]\} \times n = n$ potential outcomes (again, because the potential outcome distributions are the same in the study and target populations due to random selection, exchangeability for internal validity identifies the average potential outcomes in the unselected subjects as well). The width of the bounds attributable to not assuming exchangeability for internal validity is thus $(n/2n) \times 2 = 1$ when random selection from the target is used. In Table 1, we present the widths of the bounds attributable to uncertainty regarding exchangeability for internal and external validity in each setting explored in this paper.

These concepts are illustrated by the numerical example presented in Figure 2. Consider a target population of 100 subjects, of whom 20 are selected (not at random). Among those selected, 5 are treated (not at random). There are 180 unidentified potential outcomes remaining after causal consistency is applied. Without any exchangeability assumptions, the bounds have width $(180/200) \times 2 = 1.8$. Assuming exchangeability for internal validity allows us to identify the averages of the 15 potential outcomes under treatment for persons who are untreated and of the 5 potential outcomes under no treatment for persons who are treated, for a total of 20 potential outcomes (note that $P(S = 1) = (20/200) \times 2 = 0.2$ is the width of the bounds attributable to not assuming exchangeability for internal validity—that is, the portion of the total width of the bounds (1.8) attributable to internal validity is equal to the probability of being in the study population). Next, exchangeability for external validity allows us to identify both average potential outcomes (under treatment and no treatment) for the 80 unselected subjects, for a total of 160 potential outcomes (note that the width of the bounds attributable to not assuming exchangeability for external validity is thus $2 \times (1 - P(S = 1)) = (160/200) \times 2 = 1.6$). In this example, as in any example where $P(S = 1) < 2/3$, we can conclude that the assumption of exchangeability for external validity carries more information than the assumption of exchangeability for internal validity.

Table 1. Contribution of Exchangeability for Internal and External Validity to the Width of Bounds for the Causal Effect of a Binary Treatment on a Binary Outcome Under 4 Different Study Designs^a

Randomized?		Contribution of Exchangeability		
Treatment	Selection	Internal Validity	External Validity	Total
No	No	$P(S = 1)$	$2 \times (1 - P(S = 1))$	$2 - P(S = 1)$
Yes	No	0	$2 \times (1 - P(S = 1))$	$2 \times (1 - P(S = 1))$
No	Yes	1	0	1
Yes	Yes	0	0	0

^a $S = 1$ if a subject is selected into the study population (0 otherwise).

Figure 2. Numerical example demonstrating the relative contributions of exchangeability assumptions for internal and external validity to the width of the bounds for the causal effect of a binary treatment on a binary outcome. A) We begin with a target population of 100 people, with 5 observed to be treated and 15 observed to be untreated. Vertical lines within the circles represent treatment; horizontal lines within the circles represent no treatment. Subjects below the heavy black horizontal line are members of the study population. B) Each subject has 2 potential outcomes, 1 under treatment and 1 under no treatment. Causal consistency means that a subject's potential outcome under his or her treatment exposure is equal to his/her observed outcome. In this and each subsequent panel, the circles within dotted boxes represent potential outcomes identified by the corresponding assumptions. C) Exchangeability for internal validity identifies the unobserved potential outcomes under treatment for the untreated subjects and under no treatment for the treated subjects in the study population. D) Finally, exchangeability for external validity identifies the potential outcomes under both treatment and no treatment for the unselected subjects.

DISCUSSION

In this paper, we have described bounds that do not require exchangeability assumptions for the causal effect of a binary treatment on a binary outcome in 4 common study settings. We showed that when neither treatment nor selection is randomized, the bounds have a width of $2 - P(S = 1)$. Next, we showed that when treatment is randomized but subjects are not randomly selected from the target population, the bounds have a width of $2 \times (1 - P(S = 1))$. On the other hand, when treatment is not randomized but subjects are randomly selected from the target population, the bounds always have a constant width of 1. Finally, when both treatment and selection are randomized, the causal effect of interest is point-identified. These results are summarized in Table 1. Our derivations can be extended to the situation of a bounded continuous outcome, in which case the widths of the bounds are rescaled by a factor proportional to the range of the outcome.

Intuitively, the relative strength of the exchangeability assumptions for internal and external validity should depend on the proportion of the target population that is selected into the study. When only a small fraction of the target population is selected, as is typical, the internal validity assumption only applies to that small fraction of subjects, and thus carries much less information than the external validity assumptions. On the other hand, when most of the target population is selected into the study, the assumptions needed to extrapolate the findings to the unselected subjects are relatively weak. In the case where none of the target population is selected, such as when transporting an effect estimate from one population to another, the bounds have width $2 \times (1 - P(S = 1)) = 2 \times 1 = 2$, which covers the entire possible range of the risk difference. Exchangeability assumptions are thus needed to provide any information about effects in completely external populations.

Note that although we report the width of the bounds, we do not report the probability distribution of the effect estimate. Depending on prior knowledge of the treatment assignment and study selection processes, the width of the bounds (19) and thus the relative strength of the assumptions for internal and external validity may differ from those presented here. For instance, if it is known that subjects were approximately randomly selected from the target population, perhaps conditional on covariates, it may be more plausible that the assumptions for external validity hold, and thus those assumptions may be weaker than implied by the bounds that do not require any exchangeability assumptions. Additionally, prior knowledge regarding the mechanism of the treatment's effect on the outcome may obviate concerns about external validity if it is reasonable to assume that the effect is homogenous across populations on the scale of interest. Finally, in many situations, knowledge of the causal effect in the study population is of direct interest to investigators. In this case, the effect estimated in the randomized study is point-identified, while that from the observational study can only be set-identified with bounds of width 1 without further assumptions.

As we stated in the Introduction, our results do not account for random error, though methods have been developed for constructing valid confidence intervals for set-identified parameters (20, 21). We also note that the bounds we derived will differ for nonbinary treatments. Similar results for ratio measures may be of interest to investigators. Unfortunately, because the range for ratio measures is infinitely wide, the approach we took to quantifying the relative contribution of each assumption to the width of the bounds cannot be applied. However, the general intuition obtained by determining which potential outcomes are identified by which assumption is independent of effect measure.

Our results have important implications for choosing an appropriate study design and properly considering strategies for handling threats to internal and external validity. When

estimating an effect in a specified target population, if faced with the choice between a randomized trial without random selection from the target population and an observational study with random selection from the target population, investigators should consider the strength of the assumptions needed for each to be valid. In the former case, the assumptions correspond to bounds of width $2 \times (1 - P(S = 1))$, whereas in the latter the assumptions correspond to bounds of width 1. Therefore, if less than half of the target population is going to be selected into the study, the assumptions needed for the trial results to be valid (that is, of exchangeability between the study and target populations) are stronger than the assumptions for the results of the observational study to be valid (that is, of exchangeability across treatment arms within the study population). Similarly, when considering the results from an observational study without random selection from the target population, researchers should consider the fact that the assumptions needed for external validity are stronger than those for internal validity when less than two-thirds of the target population is selected. This intuition will be helpful in the design, analysis, and interpretation of studies and will aid in determining the best methods for estimating causal effects of interest.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Alexander Breskin, Daniel Westreich, Stephen R. Cole, Jessie K. Edwards).

This work was funded by National Institutes of Health grants DP2HD084070 and K01AI125087.

Conflict of interest: none declared.

REFERENCES

- Westreich D, Edwards JK, Lesko CR, et al. Target validity and the hierarchy of study designs. *Am J Epidemiol*. 2019;188(2):438–443.
- Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688–701.
- Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health*. 2004;58(4):265–271.
- van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):1043.
- Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–960.
- Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60(7):578–586.
- Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37–48.
- Robins JM. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Service Research Methodology: A Focus on AIDS*. Washington, DC: US Public Health Service; 1989:113–159.
- Manski CF. Nonparametric bounds on treatment effects. *Am Econ Rev*. 1990;80(2):319–323.
- Cole SR, Hudgens MG, Edwards JK. A fundamental equivalence between randomized experiments and observational studies. *Epidemiol Method*. 2016;5(1):113–117.
- Westreich D, Edwards JK, Rogawski ET, et al. Causal impact: epidemiological approaches for a public health of consequence. *Am J Public Health*. 2016;106(6):1011–1012.
- Manski CF. Identification problems in the social sciences. *Sociol Methodol*. 1993;23(1993):1–56.
- Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. *J Am Stat Assoc*. 1997;92(439):1171–1176.
- Swanson SA, Holme Ø, Løberg M, et al. Bounding the per-protocol effect in randomized trials: an application to colorectal cancer screening. *Trials*. 2015;16:Article 541.
- Lesko CR, Buchanan AL, Westreich D, et al. Generalizing study results: a potential outcomes perspective. *Epidemiology*. 2017;28(4):553–561.
- Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20(1):3–5.
- VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20(6):880–883.
- Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*. 2010;21(6):872–875.
- Lee DS. Training, wages, and sample selection: estimating sharp bounds on treatment effects. *Rev Econ Stud*. 2009;76(3):1071–1102.
- Imbens GW, Manski CF. Confidence intervals for partially identified parameters. *Econometrica*. 2004;72(6):1845–1857.
- Vansteelandt S, Goetghebeur E, Kenward MG, et al. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat Sin*. 2006;16(3):953–979.