Published in final edited form as:
Epidemiology. 2017 July ; 28(4): 562–566. doi:10.1097/EDE.0000000000000665.

“All generalizations are dangerous, even this one.” - Alexandre Dumas

Laura B. Balzer[1]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health

In public health and medicine, there is a tension between internal and external validity.[1–12] Many interventions are known to be efficacious at the individual level, but less is known about 1) their impact once scaled to a population level, 2) effect modification by both measured and unmeasured factors, and 3) which intervention components should be implemented universally and which adapted to local context. The field of HIV prevention and treatment provides an illustrative example. Randomized trials and observational studies have shown that immediate initiation of antiretroviral therapy for an HIV-positive individual improves his/her health and prevents transmission between couples and from mothers to children.[13–16] Four community randomized trials aim to examine the impact of ‘Universal Test-and-Treat’ (population-wide HIV testing with immediate antiretroviral therapy initiation for all HIV-positive persons) on HIV incidence in several countries in Eastern and Southern Africa.[17–20] These trials are pragmatic in that they aim to learn about effectiveness and implementation in real world conditions. Nonetheless, the specific components of their interventions, their implementation, and their impact are expected to vary *within* and *across* trials. Given promising interim results,[21] open questions remain about nation-wide rollout and the heterogeneity in expected impact.

In this issue of *Epidemiology*, Lesko *et al.*[12] highlight the distinction between estimating the effect for the study units and the effect for the target population. Concretely, the sample average treatment effect[22–23] is the mean difference in the counterfactual (potential) outcomes for the enrolled units, while the population average treatment effect is the expected difference in the counterfactual (potential) outcomes for the population from which the study units were selected. It is worth emphasizing that sample and population effects are fundamentally different causal parameters - even when the units are drawn as a simple random sample from the target population, there is only one version of the treatment, and there is no interference.[1–2,11] In other words, if all the assumptions for both internal and external validity held, the sample and population effects are likely to be different.

Corresponding author: Laura B. Balzer, Harvard T.H. Chan School of Public Health, Department of Biostatistics, 655 Huntington Ave., Boston, MA 02115, lbbalzer@hsph.harvard.edu, 203-558-3804.

Conflicts of interest: None declared

Computing code and data: https://github.com/LauraBalzer/On-Generalizability

ABOUT THE AUTHOR

LAURA B. BALZER is a post-doctoral fellow in Biostatistics at the Harvard T.H. School of Public Health. She is a methodologist with substantive interests in global health, community-based participatory research, and social determinants of health. Laura is an expert causal inference and semi-parametric estimation. With Maya Petersen, she was awarded the 2014 ASA’s Causality in Statistics Education Award.

Consider the simulation study conducted by Lesko *et al.*,[12] hereafter called "the authors". There is one value of the population average treatment effect, calculated analytically or by Monte Carlo simulations as −5.5%. In contrast, the sample average treatment effect is a data-adaptive parameter; with each new selection of study units, a new value of the sample effect is obtained. To illustrate this point, we replicate the authors' simulation 5000 times. For increasing enrollment sizes, we draw a simple random sample of n units and calculate the sample effect as the average difference in the counterfactual outcomes for the enrolled units (R code in Appendix C). The resulting minimum, median, and maximum values of the sample effect and its variability from our study are shown in Table 1. For the smallest size of 100, the true value of the sample effect ranges from −14.8% to 9.2% with a median value of −5.8%. In some studies the intervention is highly protective and in others quite harmful. Our simple simulation highlights the potential dangers of immediately generalizing the sample effect to the population level, even when the conditions for internal and external validity hold.

The sample effect is also an appealing causal parameter, but has received less attention in public health literature. As discussed by the authors, we commonly assume the existence of a real or hypothetical target population from which the study units were selected and about which we wish to make inferences. Concretely defining this population is challenging. In contrast, the sample effect avoids all assumptions about a "vaguely defined super-population of study units"[7] and is simply the intervention effect for units at hand. In the SEARCH trial, for example, the sample effect corresponds to the average difference in the counterfactual cumulative HIV incidence under the test-and-treat strategy and under the standard of care for the $n=32$ study communities.[11] In this example, the sample effect captures the impact for approximately 320,000 people living in rural Uganda and Kenya. As Lesko *et al.*[12] and others[1–11] discuss, generalizing this intervention effect to a wider population (e.g. all of Uganda and Kenya) or transporting it to a different population (e.g. Boston or San Francisco) requires additional assumptions and distinct estimators. Finally, the sample effect will be estimated with at least as much precision as the population effect,[1,22–23] especially under pair-matching.[11,24] In particular, if there is heterogeneity in the intervention effect by measured or unmeasured factors, a given study will have more power to detect a sample than a population effect. In other words, the price for generalizing the sample to the population is higher variance. Altogether, the interpretability, relevance, and increased precision from specifying the sample average treatment effect as the target of inference make this causal parameter an appealing alternative to the population average treatment effect.

The authors' presentation is focused on the enrollment (sampling) mechanism, and their estimators are derived under the potential outcomes framework.[1–4,7,9,22–26] An alternative would be to consider the structural causal model of Pearl.[27] Recall the authors' notation with W as the set of characteristics influencing enrollment, S as indicator of being selected into the study, A as an indicator of receiving the exposure, and Y as the outcome. Specifically, we consider a binary exposure with A=1 for the intervention and A=0 for the control. For simplicity we define the exposure and outcome (A, Y) to be zero for units not selected ($S=0$). Then the following causal model would describe a study (observational or randomized) wherein units are enrolled as a function of baseline covariates and the exposure is assigned as a function of baseline covariates and enrollment:

$$W=f_W(U_W)$$

$$S=f_S\,(W,U_S)$$

$$A=f_A(W,S,U_A)$$

$$Y=f_Y\,(W,S,A,U_Y)$$

Here $(U_W,\ U_S,U_A,\ U_Y)$ denote the corresponding set of background or unmeasured factors that have some joint distribution $\mathbb{P}_U$. This approach for representing selection into the study (a threat to external validity) could easily be extended to include missing data after enrollment (a threat to internal validity) as well as post-enrollment covariates that influence the exposure assignment (additional confounders, another threat to internal validity) (Appendix A). Furthermore, this causal model is defined at the unit-level and implicitly assumes no interference. (This framework could also be extended to handle interference.) The authors' simulation example is one possible data generating process, compatible with this structural causal model. In a randomized trial (such as considered by the authors), the unmeasured factors contributing to the intervention assignment U_A are independent of the others and the covariates W do not impact randomization A.

We assume that the above causal model provides a description of the data generating process under existing conditions and under specific interventions.[27] (See Appendix 1 in Bareinbom and Pearl[8] for a short introduction.) Counterfactual outcomes are generated by intervening on this causal model. To define the sample effect, we intervene to set the exposure $A=a$ to generate the counterfactual outcome $Y_i(a)$: the outcome that would have been observed if unit i received exposure-level $A=a$. Then the sample average treatment effect (*SATE)* is defined as the average difference in these counterfactual outcomes among enrolled units ($S=1$)

$$\text{SATE}=\frac{1}{n}\sum_{i\in\{S_i=1\}}^{n} Y_i(1)-Y_i(0)$$

where n denotes the total number of units in the study. To define the population average treatment effect (*PATE*) in the context of biased sampling, we consider a hypothetical intervention to enroll the entire target population (i.e. set $S=1$) and assign the exposure $A=a$. We denote the counterfactual outcome under this joint intervention as $Y(s=1,a)$. The *PATE* is then given by the expected difference in these counterfactual outcomes across the target population of interest:

$$\text{PATE}=\mathbb{E}[\,Y(1,1)-Y(1,0)]$$

For illustration, we repeat the authors' simulation study 5000 times: 1) generate a target population of 50,000 units; 2) from that population draw a biased sample of *n=2000* units; and 3) for each sample calculate the *SATE* as the average difference in the counterfactual outcomes for the enrolled units (R code in Appendix C). As shown in Table 1, the sample effect under the biased sampling scheme ranges from −12.8% to −8.3% and is −10.4% on average. The sample effect is larger on average than the population effect (−5.5%), because units at higher risk of the outcome are selected into the study. Practically, this may suggest that instead of rolling out the intervention to the entire population, the greatest impact could be obtained by targeting the intervention to high-risk groups. Likewise, the effect heterogeneity may suggest alternative parameters of interest, such as the conditional average treatment effect, an intermediate between the sample and population effects (Appendix B).[1,28–29]

The structural causal model representation also draws a connection between the authors' assumptions and estimators for external validity and the standard machinery for controlling for confounding, selection bias, and/or unrepresentative sampling.[3–4,8,10,30–35] Given the sequential randomization assumption

$$Y(s=1,a)\perp S\mid W$$

$$Y(s=1,a)\perp A\mid S=1,\ W$$

and the corresponding positivity assumptions, we have the G-computation identifiability result:[32]

$$\text{GComp.}=\sum_{w}[\mathbb{E}(Y|A=1,S=1,W=w)-\mathbb{E}(Y|A=0,S=1,W=w)]\mathbb{P}(W=w)$$

This estimand is written equivalently in inverse probability weighting (*IPW*) form as

$$\text{IPW}=\mathbb{E}\left[\left(\frac{\mathbb{I}(A=1,S=1)}{\mathbb{P}(A=1,S=1|W)}-\frac{\mathbb{I}(A=0,S=1)}{\mathbb{P}(A=0,S=1|W)}\right)Y\right]$$

where the weights could be factorized into a product of propensity score $\mathbb{P}(A=1|S=1,W)$ and selection mechanism $\mathbb{P}(S=1|W)$.[4,30] While stabilized weights are also possible,[3] the above estimands showcase the equivalence when non-parametric estimators are used for the outcome regression and selection/exposure mechanisms in both observational and trial settings.

For our simulation study, we implemented the corresponding G-computation (a.k.a. standardization)[36–37] and IPW estimators for the population effect. For comparison we also implemented the unadjusted estimator, as the difference in the mean outcomes between enrolled treated units and enrolled control units. The results of 5000 repetitions are shown in Table 2 (R code in Appendix C). As expected, the unadjusted estimator is unbiased for the sample effect, but exhibits substantial bias when the target of inference is the population effect. Also as expected, the G-computation and IPW estimators are able to correct for the biased sampling scheme and are identical. (The algorithms are equivalent when non-parametric estimators are used for the outcome regression and the selection/exposure mechanism.)

The structural causal model representation also emphasizes that a rich toolkit of estimators could be used to correct for biased sampling, which is presented by the authors as a threat to external validity. The non-parametric estimators, implemented by the authors and here, will break down when many covariates or a single continuous covariate influence unit selection (and/or the exposure mechanism in an observational setting). As an alternative, we could immediately implement augmented inverse probability weighting or targeted maximum likelihood estimation.[35,38–39] These methods are double robust and can incorporate data-adaptive (machine learning) algorithms to relax parametric modeling assumptions, while retaining valid statistical inference.

An important open question, not addressed by Lesko *et al.*[12] nor in this commentary, is generalizability and transportability when the intervention (or its specific components) must be adapted to local context. We should be wary assuming the sample effect is immediately generalizable to the population. We should also be wary of assuming that a one-size-fits-all intervention is best.

Acknowledgments

We thank Drs. Maya Petersen and Victor DeGruttola for their expert advice.

Financial support: This work was supported by grants R37AI051164 and U01AI099959 from the National Institute of Allergy & Infectious Diseases (NIAID of the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Appendix A

Let Z denote the set of post-enrollment characteristics influencing exposure assignment and Δ be an indicator that a unit has its outcome measured (i.e. is not loss to follow-up). For simplicity, we define the post-enrollment characteristics, the exposure, the missing data indicator, and the outcome (Z, A, Δ, Y) equal to zero for units not enrolled (S=0). Then the following structural causal model would describe a study (observational or randomized) wherein units are enrolled as a function of baseline covariates, the exposure is rolled out as a function of baseline and post-enrollment characteristics, and missingness on the outcome is not random:

$$W = f_W(U_W)$$

$$S = f_S\left(W, U_S\right)$$

$$Z = f_Z\left(W, S, U_Z\right)$$

$$A = f_A\left(W, S, Z, U_A\right)$$

$$\Delta = f_\Delta\left(W, S, Z, A, U_\Delta\right)$$

$$Y = f_Y\left(W, S, Z,\ A, \Delta, U_Y\right)$$

Here $\left(U_W,\ U_S, U_Z, U_A,\ U_\Delta, U_Y\right)\ \sim \mathbb{P}_U$ denote the corresponding set of background or unmeasured factors with some distribution. Let $Y\left(s=1, a, \delta=1\right)$ denote the counterfactual outcome for a given unit under a hypothetical intervention to ensure its enrolled (i.e. set *S=1)*, assign the exposure *A=a*, and ensure its outcome is measured (i.e. set $\Delta=1)$. Then the PATE is defined as

$$\mathbb{E}[\left(Y\left(1,1,1\right) - Y\left(1,0,1\right)\right]$$

Under the sequential randomization and positivity assumptions,[32] the corresponding statistical estimand could be estimated with a variety of methods, including longitudinal parametric G-computation,[40] longitudinal inverse probability weighting[33,41], and longitudinal targeted maximum likelihood estimation.[35,42]

Appendix B

In 2002 Abadie and Imbens[28] proposed the conditional average treatment effect as

$$\text{CATE} = \frac{1}{n}\sum_{i \in \{S_i = 1\}}^{n} \mathbb{E}[Y_i\left(1\right) - Y_i\left(0\right) | W_i]$$

where i indexes the *n=2000* units selected for the study. The conditional effect is interpreted as the average intervention effect given the covariates of the study units and is equal to −10.4% under this biased sampling scheme:

$$\frac{1}{2000}\left[320 \times \left(-0.05\right) + 480 \times \left(-0.05\right) + 480 \times \left(-0.05\right) + 720 \times \left(-0.05 - .15\right)\right] = -10.4\%.$$

Appendix C

Simulation studies were conducted in R-3.3.2.[43] Full R code and the resulting data set are available at https://github.com/LauraBalzer/On-Generalizability.

References

1. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat. 2004; 86(1):4–29.
2. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc Ser A. 2008; 171(Part 2):481–502.
3. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 Trial. Am J Epidemiol. 2010; 172(1):107–115. [PubMed: 20547574]
4. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc Ser A. 2011; 174(Part 2):369–386.
5. Rothman KJ, Gallacher JEJ, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013; 42:1012–1014. [PubMed: 24062287]
6. Elwood JW. On representativeness [Commentary]. Int J Epidemiol. 2013; 42:104–1015.
7. Schochet P. Estimators for clustered education RCTs using the Neyman model for causal inference. J Educ Behav Stat. 2013; 38(3):219–238.
8. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. J Causal Inference. 2013; 1(1):107–134.
9. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. J R Stat Soc Ser A. Jan; 2015 178(3):757–778.
10. Pearl J. Generalizing experimental findings. J Causal Inference. 2015; 3(2):259–266.
11. Balzer LB, Petersen ML, van der Laan MJ, the SEARCH Collaboration. Targeted estimation and inference of the sample average treatment effect in trials with and without pair-matching. Stat Med. 2016; 35(21):3717–3732. [PubMed: 27087478]
12. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology. 2017
13. The TEMPRANO ANRS 12136 Study Group. A trial of early antiretrovirals and isoniazid preventive therapy in Africa. N Engl J Med. 2015; 373:808–822. [PubMed: 26193126]
14. The INSIGHT START Study Group. Initiation of antiretroviral therapy in early asymptomatic HIV infection. N Engl J Med. 2015; 373(9):795–807. [PubMed: 26192873]
15. Cohen M, Chen Y, Mccauley M, Gamble T, Hosseinipour M, Kumarasamy N, et al. Final results of the HPTN 052 randomized controlled trial: antiretroviral therapy prevents HIV transmission. J Acquir Immune Defic Syndr. 2015; 18(Suppl4):20479.
16. World Health Organization. Guideline on when to start antiretrovial therapy and on pre-exposure prophylaxis for HIV. http://www.who.int/hiv/pub/guidelines/earlyrelease-arv/en/. Accessed March 12, 2017
17. French National Institute for Health and Medical Research-French National Agency for Research on AIDS and Viral Hepatitis (Inserm-ANRS). Impact of immediate versus South African recommendations guided ART initiation on HIV incidence (TasP). https://clinicaltrials.gov/ct2/show/NCT01509508. Accessed March 12, 2017
18. Centers for Disease Control and Prevention. Botswana Combination Prevention Project (BCPP). http://clinicaltrials.gov/show/NCT01965470. Accessed March 12, 2017
19. HIV Prevention Trials Network. Population Effects of Antiretroviral Therapy to Reduce HIV Transmission (PopART). http://clinicaltrials.gov/show/NCT01900977. Accessed March 12, 2017
20. University of California. San Francisco: Sustainable East Africa Research in Community Health (SEARCH); http://clinicaltrials.gov/show/NCT01864603. Accessed March 12, 2017
21. Petersen, M., Balzer, L., Kwarsiima, D., Sang, N., Chamie, G., Ayieko, J., et al. SEARCH test and treat study in Uganda and Kenya exceeds the UNAIDs 90-90-90 cascade target by achieving over

80% population-level viral suppression after 2 years. 21st International AIDS Conference; Durban, South Africa. 2016.

22. Neyman J. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990). Stat Sci. 1923; 5:465–480.
23. Rubin DB. Neyman 1923 and causal inference in experiments and observational studies [Comment]. Stat Sci. 1990; 5(4):472–480.
24. Imai K. Variance identification and efficiency analysis in randomized experiments under the matched pair design. Stat Med. 2008; 27(24):4857–4873. [PubMed: 18618425]
25. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974; 66(5):688–701.
26. Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986; 81(396):945–960.
27. Pearl, J. Causality: Models, Reasoning and Inference. Second. Vol. 2000. Cambridge University Press; New York: 2009.
28. Abadie A, Imbens G. Simple and bias-corrected matching estimators for average treatment effects. Technical Report. 2002; 283NBER technical working paper
29. Balzer LB, Petersen ML, van der Laan MJ, the SEARCH Consortium. Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. Stat Med. 2015; 34(6):999–1011. [PubMed: 25421503]
30. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952; 47:663–685.
31. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.
32. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods– application to control of the healthy worker survivor effect. Mathematical Modelling. 1986; 7:1393–1512.
33. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11(5):550–560. [PubMed: 10955408]
34. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15(5):615–625. [PubMed: 15308962]
35. van der Laan, MJ., Rose, S. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer; New York Dordrecht Heidelberg London: 2011.
36. Miettinen OS. Standardization of risk ratios. Am J Epidemiol. 1972; 96(6):383–388. [PubMed: 4643670]
37. Rothman, KJ., Greenland, S., Lash, TL. Modern Epidemiology. Lippincott Williams & Wilkins; Phildelphia: 2008.
38. Robins, JM. Robust estimation in sequentially ignorable missing data and causal inference models. 1999 Proceedings of the American Statistical Association; Alexandria, VA. American Statistical Association; 2000. p. 6-10.
39. van der Laan, MJ., Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. Springer-Verlag; New York Berlin Heidelberg: 2003.
40. Taubman SL, Robins JM, Mittleman MA, Hernan MA. Intervening on risk factors for coronary heart disease: an application of the parametric G-formula. Int J Epidemiol. 2009; 38(6):1599–1611. [PubMed: 19389875]
41. Bodnar LM, 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(10):926–934. [PubMed: 15128604]
42. Petersen ML, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan MJ. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. J Causal Inference. 2014; 2(2)
43. Core Team R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2015.

Table 1

Summary of the causal parameters (in %) over 5000 simulations under simple random sampling of size n={100, 500, 2000} and the authors' biased sampling scheme of size n=2000.

	Sample average treatment effect			
	Min	**Median**	**Max**	**Variance**
n=100 simple	−14.8	−5.8	9.2	0.10
n=500 simple	−10.2	−5.6	0.2	0.02
n=2000 simple	−8.0	−5.5	−2.9	0.006
n=2000 biased	−12.8	−10.4	−8.3	0.005

Table 2

Summary of the G-computation (Gcomp.) and inverse probability weighting (IPW) for estimation of the population effect (in %) over 5000 simulations with biased sampling of n=2000 units. Performance of the unadjusted estimator (Unadj.) for the sample and population effects are also shown. The true value of the sample effect changes with each study enrollment (Table 1); population effect is −5.5%.

	Average	Bias	Variance	MSE	Coverage[a]
Unadj. for SATE	−10.4	−0.01	0.04	0.04	95.0
Unadj. for PATE	−10.4	−4.9	0.04	0.29	30.8
Gcomp. for PATE	−5.5	−0.06	0.05	0.05	95.0
IPW for PATE	−5.5	−0.06	0.06	0.06	95.0

MSE indicates mean squared error, SATE sample average treatment effect, PATE population average treatment effect, IPW inverse probability weighting.

[a] Coverage: 95% confidence interval coverage, constructed using the true (vs. estimated) variance.