Evaluating Public Health Interventions: 3. The Two-Stage Design for Confounding Bias Reduction—Having Your Cake and Eating It Two

In public health evaluations, confounding bias in the estimate of the intervention effect will typically threaten the validity of the findings. It is a common misperception that the only way to avoid this bias is to measure detailed, high-quality data on potential confounders for every intervention participant, but this strategy for adjusting for confounding bias is often infeasible.

Rather than ignoring confounding altogether, the two-phase design and analysis—in which detailed high-quality confounding data are obtained among a small subsample—can be considered.

We describe the two-stage design and analysis approach, and illustrate its use in the evaluation of an intervention conducted in Dar es Salaam, Tanzania, of an enhanced community health worker program to improve antenatal care uptake. (*Am J Public Health.* 2016; 106:1223–1226. doi:10.2105/AJPH.2016.303250)

Donna Spiegelman, ScD, MS, Claudia L. Rivera-Rodriguez, PhD, and Sebastien Haneuse, PhD

In this commentary, we describe the two-stage design and analysis approach, and we illustrate its use in the evaluation of an intervention of an enhanced community health worker program to improve antenatal care uptake that was conducted in Dar es Salaam, Tanzania. This design and analysis strategy allows public health evaluators to use richer information in a smaller sample to adjust for confounding in observational and cluster-randomized interventions. The Familia Salama study, a public health intervention conducted in Dar es Salaam, Tanzania between 2012 and 2014, which aimed to evaluate an enhanced community health worker program, serves as our working example (Figure 1). In the Familia Salama implementation science study,[1,2] all of the nearly 60 wards in two of the three districts of Dar es Salaam were randomized to receive either standard of care or an enhanced community health worker intervention designed to improve antenatal care through education, referrals, and follow-up (X). We observed more than 200 000 pregnancies (n_1) during a nearly two-year study period for adherence to World Health Organization guidelines of attending at least four antenatal care visits during the course of their pregnancy (D).[2] As an implementation science study, we obtained the data on these 200 000 pregnancies with the existing methods for recording clinical encounters in the health system using handwritten registry books, as is common around the world.

In part because of concerns about residual confounding within wards, where the number of pregnancies within wards was as large as 5402 with a median of 934, we conducted an in-depth population survey among 2329 pregnant women (n_2) in a 2% sample of the wards included in this study. We obtained more detailed confounder data (C_2) through the survey, enhancing the limited confounder data available from the registries (C_1; district and gestational age at first antenatal care visit), including many variables related to maternal health and reproductive history, socioeconomic status, and health-related knowledge.

CONFOUNDING CONTROL

Confounding is the source of bias to which observational (i.e., nonrandomized) research is vulnerable that public health implementers, epidemiologists, and program evaluators have most grappled with to date. Control of confounding has been the primary focus of what is otherwise known as causal inference methods. Although confounding has been the primary focus of empirical bias adjustment methods, it has been argued that measurement error and misclassification are typically much more serious sources of bias in observational research, at least in some contexts.[3] However, in cluster-randomized interventions, including stepped wedge designs, "exposure" misclassification is eliminated by design. Conversely (as illustrated in the box on page 456 in my previous article in this series), residual confounding may still be operating, especially when the number of clusters is low, as is often the case.[4] Table 1 reviews the standard methods for adjusting for bias stemming from confounding.

It deserves mention that none of these methods will adequately adjust for time-varying confounders that are both a consequence of the intervention and

ABOUT THE AUTHORS

Donna Spiegelman and Claudia L. Rivera-Rodriguez are with the Departments of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA. Sebastien Haneuse is with the Department of Biostatistics, Harvard T. H. Chan School of Public Health.

Correspondence should be sent to Donna Spiegelman, Department of Epidemiology, Harvard T. H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115 (e-mail: stdls@hsph.harvard.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted April 25, 2016.

doi: 10.2105/AJPH.2016.303250

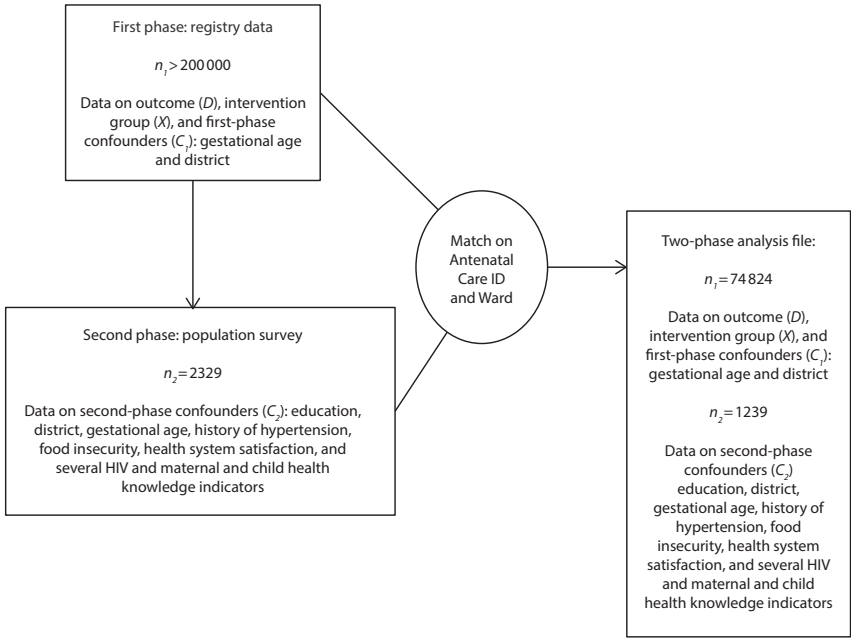

FIGURE 1—Familia Salama Two-Phase Study Design: Dar es Salaam, Tanzania, 2012–2014

a cause of disease. When time-varying interventions are studied—and when the model for the future intervention status as a function of current and past confounders can be validly estimated or when there is sufficient data to stratify on each unique joint intervention and confounder history—the methods illustrated by Hernán et al. and others can and should be used.[5] In a future article, we will further discuss these causal inference methods in relation to public health evaluations.

Standard methods for time-varying and time-invariant confounding, as well as the newer ones, require that data be available for all confounders for all participants at all relevant time points. From a design point of view, this requirement is quite daunting and forces evaluators to exclude potential confounders a priori from consideration at the design stage. Now for some good news: it turns out that evaluators do not need to choose between measuring confounders among all study participants or to ignore the confounder entirely, which are the two common choices in standard practice. Rather, two-stage designs can be used. The first stage is the standard design and includes n_1 participants with data on the outcome (D), the intervention (X), and confounders that are inexpensive to measure (C_1), if any. In the second stage, n_2 participants are sampled with probability $\Pr(I_i = 1 | D_i, X_i, C_{1i}; \phi)$ for all $i = 1, \ldots, n_2$ participants from the first stage for the assessment of data on the expensive confounders (C_2), where I_i is an indicator that takes on value 1 if participant i was sampled for the second stage and 0 otherwise. Typically, the first-stage sample size will be much greater than the second-stage sample size, that is, $n_1 \gg n_2$. In the Tanzanian Familia Salama intervention, $I_i = 1$ if participant i was included in the population survey and 0 otherwise, and $\Pr(I_i = 1) \approx 2042/200\,000 = 2\%$ overall.

TABLE 1—Standard Design and Analysis Methods to Adjust for Bias Stemming From Confounding

Design	Analysis
Randomize	Intent to treat
Match on key confounders	Matched
Restrict to a single level of key confounders	Crude
Collect data on known and suspected confounders	Multivariate models, Mantel-Haenszel methods

TWO-STAGE DESIGN

The two-stage design was first proposed in 1938,[6] was later independently reintroduced into the epidemiological and biostatistical literature by White[7] and Walker,[8] and was further developed by Cain and Breslow in 1988.[9] Although these methods were developed with observational studies in mind, they can straightforwardly be applied to randomized studies as well, as long as randomization is not clustered—as would often be the case in public health evaluations.

Rather than randomly sampling participants from the first stage for more detailed assessments of confounders or outcomes at the second stage, cost-efficient two-stage designs can be obtained by choosing the sampling fractions for selection into the second stage as a function of (D,X,C_1).[9,10] A comprehensive, nontechnical article on the two-stage designs was written by Reilly.[11] Reilly provided formulas and examples for four typical study design questions. These included designs that identified the minimum variance—or most cost-efficient—second-stage sampling fractions for estimating the intervention effect over the unique combinations of the strata formed by (D,X,C_1). These designs are relevant when the first-stage data are freely available, for example, from an electronic medical records database or disease registry, and there is a fixed budget for sampling a specified number of participants at the second stage to obtain more detailed confounder data through personal interviews,

medical records review, or laboratory measurements.

In addition, Reilly provided public freeware in Stata, R, and Splus to implement these design calculations (see http://www.mep.ki.se/%7Emarrei/software for links and a user manual). Breslow and Cain[12] suggested that one will typically approach the most efficient design by choosing second-stage sampling probabilities that lead to the inclusion of approximately the same numbers of first-stage participants among each unique combination of (C_1, X,D) that occur in the first stage. Two-stage designs do not require sampling on outcome, and efficiency can still be gained by sampling into the second stage jointly on (C_1,X). Finally, sampling into the second stage can be entirely at random, as in the Familia Salama study.

The box below provides further details on the analysis of two-stage designs.

USAGE IN PUBLIC HEALTH

Have these methods for design and analysis been used by public health researchers to improve control for confounding? Hardly! A reverse citation search of the seminal 1988 article by Cain and Breslow[9] found that among 72 citations, 5 linked to likely substantive applications of the methodology (e.g., Gilliland et al.[13] and Morabia et al.[14]) and none to public health evaluations, whereas among the 203 subsequent citations of the 1988 companion article by Breslow and Cain,[12] a reverse citation search turned up an additional two relevant substantive articles that clearly used this methodology (e.g., Engels et al.[15]).

The generalized estimating equations methods for analysis of two-stage designs fared even worse in terms of uptake. Of 136 citations to Reilly and Pepe,[16] one clearly linked to a substantive article[17]; similarly, one[18] for the 112 citations linked to Flanders and Greenland.[19] Walker's seminal article[8] appears to have been applied twice (e.g., Hsieh et al.[20]) among 41 citations, both of which are coauthored by Walker himself, whereas White's seminal article[7] has received 122 citations with 3 being clearly substantive (e.g., Ritz et al.[21] and Strauss et al.[22]). One of these describes the design of the Healthy Communities Study, which assesses the impact of characteristics of community programs and policies targeting childhood obesity and children's body mass index, diet, and physical activity.[22] This is the only citation we have been able to find that shows the application of two-stage design and analyses methods to an implementation science project and its evaluation.

Why such a gap between the development of this promising methodology and its widespread application? One barrier could be the perception of the unavailability of user-friendly software, which is ideally implemented in the packages most commonly used by practitioners. In fact, packages for the analysis of two-stage designs are available for SAS users[23] and R users.[24]

Another possible reason for the poor uptake of two-stage design and analysis methodology in public health evaluation is that this now quite large body of methodologic work cannot properly account for clustering, for example by facility, village, or other geographic or social units. Our new work is filling this gap (Rivera-Rodriguez CL, Spiegelman D, Haneuse S, unpublished data, 2016). In a synthetic illustrative example applied to a routine monitoring and evaluation setting in Malawi, Haneuse et al.[25] demonstrated that data aggregated at the clinic-quarter level —as is common practice, for example, in the President's Emergency Plan for AIDS Relief monitoring and evaluation programs—provided biased estimates of the differences by gender, age, and private and public clinic in routine clinical quality indicators. On the other hand, a synthetic two-stage design sampling 5000 of the 82 997 clinics provided effect estimates nearly identical to the full data with little loss in precision.[25] This article serves as a powerful illustration of the depth of information that two-stage designs can provide.

FAMILIA SALAMA

To further illustrate these new methods, we provide some preliminary results from the Familia Salama study. We calculated the intervention effect in the first stage alone, in which it is possible to adjust for residual confounding within ward for two potential confounders only (C_1); the fully adjusted but small second stage; and then with the two-stage analyses using the two approaches we have discussed (Table 2). Before the availability of these new methods, the investigators, with whom we work as the statisticians on the project team, had planned to completely discard the data on more than 200 000 pregnancies and use the second-stage data—the population survey—only ($n_2 = 2042$). In the two-stage analysis, we were able to match nearly 75 000 first-stage participants from the clinic registries to 1239 population survey participants (Figure 1). The estimated intervention effects, their

NOTE ON THE ANALYSIS OF TWO-STAGE DESIGNS

Once a two-stage design has been conducted, as long as the probability of being sampled into the second stage depended multiplicatively on the outcome and intervention because of the first-stage confounders, that is, if the sampling probabilities into the second stage were a product of two second-stage sampling probabilities, $\Pr(I_i = 1|D_i,X_i,C_{1i};\phi) = \Pr(I_i = 1|D_i,C_{1i};\phi_0)\Pr(I_i = 1|D_i,C_{1i};\phi_E)$ that could be fit, for example, by a logistic regression model of this form

$$logit\,[\Pr(I_i = 1|D_i, X_i, C_{1i};\phi)] = \phi_0 + \phi_1 D_i + \phi_2 X_i + \phi_3^T C_1,$$

with no interaction between D_i and X_i. As may often be the case when the odds ratio is the parameter of interest, the sampling fractions can be ignored in the subsequent analysis although the efficiency of the intervention effect estimate may be less than what could otherwise be obtained using the analytic approaches. Otherwise, the valid and efficient analytic method will be different from usual approaches. Maximum likelihood methods will often be the most efficient but require the development of custom software for each new study. Pseudolikelihoods can be maximized, and other sorts of consistent estimating equations can be solved for the intervention effect, which give consistent but not fully efficient estimates of the intervention effect (e.g., Breslow and Cain,[12] Reilly and Pepe,[16] and Flanders and Greenland[19]).

TABLE 2—Preliminary Results: The Effect of an Enhanced Community Health Worker Intervention on Increased Compliance With World Health Organization Recommendations for Antenatal Care: Familia Salama Study, Dar es Salaam, 2013–2014

		Unadjusted		Multivariate-Adjusted[a]	
	No.	OR (95% CI)	*P*	OR (95% CI)	*P*
First stage (registry)	74 824	2.51 (1.72, 3.68)	≤ .001	NA	
Second stage (survey)	1 239	1.66 (1.03, 2.67)	.04	1.97 (1.13, 3.42)	.02
2-stage weighted GEE[16,19]	74 824	[b]		2.88 (1.75, 4.75)	≤ .001
2-stage pseudolikelihood[9,12]	74 824	[b]		2.93 (2.04, 4.19)	≤ .001

Note. CI = confidence interval; GEE = generalized estimating equations; OR = odds ratio.

[a]Adjusted for education, district, gestational age, history of hypertension, food insecurity, health system satisfaction, and several HIV and maternal and child health knowledge indicators.

[b]Same as first stage.

95% confidence intervals and the *P* values for the test of the null hypothesis are given in Table 2.

These results suggest that that in the "intent to treat" analysis, the enhanced community health worker intervention significantly increased the odds of compliance with World Health Organization antenatal care visits by 2.5-fold. They also suggest that the effect estimate increased to nearly 3-fold after extensive adjustment for residual confounding within and between wards. This ruled out confounding as a source of bias through the use of a 2% second-stage subsample, with no loss of power whatsoever and at less than 2% additional cost.

RECOMMENDATIONS

We recommend that public health evaluators consider two-stage designs and analyses whenever the primary data source for the evaluation is thin on high-quality data on potential confounders. Evaluators will also find this design to be of great utility in devising an evaluation with causal inferential potential within a reasonable, usually slim budget. We hope we have convinced you that, with the addition of the second stage, partaking from the causal cake will not be restricted to high-budget biomedical clinical trialists! AJPH

CONTRIBUTORS

D. Spiegelman wrote the article. D. Spiegelman and S. Haneuse contributed to data analysis. C.L. Rivera-Rodriguez analyzed the data. C.L. Rivera-Rodriguez and S. Haneuse contributed to writing the article.

ACKNOWLEDGMENTS

The authors thank the investigators and participants of the Familia Salama Study, Management and Development for Health, Dar es Salaam, Tanzania.

HUMAN PARTICIPANT PROTECTION

Familia Salama was approved by the National Institutes of Medical Research in Tanzania and the Harvard T.H. Chan School of Public institutional review board.

REFERENCES

1. Lema IA, Sando D, Magesa L, et al. Community health workers to improve antenatal care and PMTCT uptake in Dar es Salaam, Tanzania: a quantitative performance evaluation. *J Acquir Immune Defic Syndr*. 2014;67(suppl 4):S195–S201.

2. Sando D, Geldsetzer P, Magesa L, et al. Evaluation of a community health worker intervention and the World Health Organization's Option B versus Option A to improve antenatal care and PMTCT outcomes in Dar es Salaam, Tanzania: study protocol for a cluster-randomized controlled health systems implementation trial. *Trials*. 2014;15:359.

3. Blair A, Stewart P, Lubin JH, Forastiere F. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *Am J Ind Med*. 2007;50(3):199–207.

4. Spiegelman D. Evaluating public health interventions: 2. Stepping up to routine public health evaluation with the stepped wedge design. *Am J Public Health*. 2016;106(3):453–457.

5. 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(5):561–570.

6. Neyman J. Contribution to the theory of sampling human populations. *J Am Stat Assoc*. 1938;33(201):101–116.

7. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol*. 1982;115(1):119–128.

8. Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*. 1982;38(4):1025–1032.

9. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *Am J Epidemiol*. 1988;128(6):1198–1206.

10. Tosteson TD, Ware JH. Designing a logistic regression study using surrogate measures for exposure and outcome. *Biometrika*. 1990;77:11–21.

11. Reilly M. Optimal sampling strategies for two-stage studies. *Am J Epidemiol*. 1996;143(1):92–100.

12. Breslow NE, Cain KC. Logistic-regression for 2-stage case–control data. *Biometrika*. 1988;75(1):11–20.

13. Gilliland FD, Hunt WC, Baumgartner KB, et al. Reproductive risk factors for breast cancer in Hispanic and non-Hispanic White women—the New Mexico Women's Health Study. *Am J Epidemiol*. 1998;148(7):683–692.

14. Morabia A, Bernstein M, Heritier S, Khatchatrian N. Relation of breast cancer with passive and active exposure to tobacco smoke. *Am J Epidemiol*. 1996;143(9):918–928.

15. Engels EA, Chen J, Viscidi RP, et al. Poliovirus vaccination during pregnancy, maternal seroconversion to simian virus 40, and risk of childhood cancer. *Am J Epidemiol*. 2004;160(4):306–316.

16. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*. 1995;82(2):299–314.

17. Surkan PJ, Hsieh CC, Johansson ALV, Dickman PW, Cnattingius S. Reasons for increasing trends in large for gestational age births. *Obstet Gynecol*. 2004;104(4):720–726.

18. Behr S, Schill W, Pigeot I. Does additional confounder information alter the estimated risk of bleeding associated with phenprocoumon use—results of a two-phase study. *Pharmacoepidemiol Drug Saf*. 2012;21(5):535–545.

19. Flanders WD, Greenland S. Analytic methods for 2-stage case–control studies and other stratified designs. *Stat Med*. 1991;10(5):739–747.

20. Hsieh CC, Crosson AW, Walker AM, Trapido EJ, Macmahon B. Oral contraceptive use and fibrocystic breast disease of different histologic classifications. *J Natl Cancer Inst*. 1984;72(2):285–290.

21. Ritz B, Wilhelm M, Hoggatt KJ, Ghosh JKC. Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles. *Am J Epidemiol*. 2007;166(9):1045–1052.

22. Strauss WJ, Sroka CJ, Frongillo EA, et al. Statistical design features of the healthy communities study. *Am J Prev Med*. 2015;49(4):624–630.

23. Schill W, Enders D, Drescher K. A SAS package for logistic two-phase studies. *J Stat Softw*. 2014;57(9):1–13.

24. Haneuse S, Saegusa T, Lumley T. osDesign: an R package for the analysis, evaluation, and design of two-phase and case–control studies. *J Stat Softw*. 2011;43(11):1–29.

25. Haneuse S, Hedt-Gauthier B, Chimbwandira F, Makombe S, Tenthani L, Jahn A. Strategies for monitoring and evaluation of resource-limited national antiretroviral therapy programs: the two-phase design. *BMC Med Res Methodol*. 2015;15:31.