American Journal of Epidemiology
© The Author 2010. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 171, No. 6
DOI: 10.1093/aje/kwp436
Advance Access publication:
February 5, 2010

Invited Commentary

Invited Commentary: Positivity in Practice

Daniel Westreich* and Stephen R. Cole

* Correspondence to Dr. Daniel Westreich, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, McGavran-Greenberg Hall, CB 7435, Chapel Hill, NC 27599-7435 (e-mail: djw@unc.edu).

Initially submitted October 22, 2009; accepted for publication October 28, 2009.

Positivity, or the experimental treatment assignment assumption, requires that there be both exposed and unexposed participants at every combination of the values of the observed confounders in the population under study. Positivity is essential for inference but is often overlooked in practice by epidemiologists. This issue of the *Journal* includes 2 articles featuring discussions related to positivity. Here the authors define positivity, distinguish between deterministic and random positivity, and discuss the 2 relevant papers in this issue. In addition, the commentators illustrate positivity in simple 2 × 2 tables, as well as detail some ways in which epidemiologists may examine their data for nonpositivity and deal with violations of positivity in practice.

causality; causal inference; confounding factors (epidemiology); epidemiologic methods; exchangeability; identifiability; multilevel analysis; propensity score

Positivity (1, 2), or the experimental treatment assignment assumption (3), is a necessary assumption for causal inference in observational data, along with consistency (4), exchangeability (i.e., no unmeasured confounding and no selection bias), no measurement error, no interference, and correct model specification. Positivity requires that there be both exposed and unexposed participants at every combination of the values of the observed confounder(s) in the population under study. Formally, for a discrete-valued exposure X and an arbitrary confounder vector $\mathbf{Z}$, if $f(\mathbf{Z}) \neq 0$ then $\Pr(X = x|\ \mathbf{Z}) > 0$ for all $x \in X$, where $f(\cdot)$ is the probability density function. The prior definition extends to time-varying exposures and confounders (2). Positivity is present by design in randomized controlled trials, in which every trial participant has a known probability of each treatment under study (often one-half).

In observational data, violations of positivity may be deterministic or random. A deterministic violation is one in which participants at 1 or more levels of the confounders cannot receive at least 1 level of the exposure. For example, because men lack a uterus, they cannot receive a hysterectomy in a study of the effects of hysterectomy on mortality. For other examples of deterministic nonpositivity, see Cole and Hernán (2). In contrast, random violations of positivity occur when, at 1 or more levels of the confounders, no one happens to be observed at 1 or more levels of the exposure. Random nonpositivity may be further classified by whether or not the nonpositivity is "surrounded" by regions of positivity. For example, consider that older age is associated with aspirin use and myocardial infarction; it is plausible that a small observational study of daily aspirin intake for prevention of myocardial infarction might show the exposure pattern depicted in Table 1, section A. By chance, no one aged 31–35 years was exposed. Also by chance, no one aged 41–45 years was exposed. Both are examples of random nonpositivity, but an investigator might be more comfortable interpolating to ages 41–45 years than extrapolating to ages 31–35 years.

The issue of positivity is clearly important when attempting to make inferences from observational data. However, the assumption is rarely assessed in biomedical research (3). In this issue of the *Journal*, 2 articles related to birth outcomes feature relevant discussions (5, 6). In the former, Cheng et al. (5) examine the association between fetal position and perinatal outcomes. They observe that an analysis which matches on the propensity score excludes persons who do not match and thus avoids extrapolation, albeit at the cost of redefining the estimand to be the effect in persons who match rather than the total study population (5). In the latter, Messer et al. (6) examine the independent effects of socioeconomic status and race on preterm delivery in 2 North Carolina counties. Messer et al. find that in these data there were (almost) no poor, all-white census tracts and (almost) no rich, all-black census tracts. A naive multilevel

Table 1. Hypothetical Data for Exposure to Daily Aspirin Use by Age in A) 5-Year Age Categories and B) 10-Year Age Categories, Showing 2 Types of Random Nonpositivity

Exposure Status	Age Group, years					
A)	31–35	36–40	41–45	46–50	51–55	56–60
Exposed	0	2	0	3	5	7
Unexposed	9	7	9	6	4	2
B)	31–40		41–50		51–60	
Exposed	2		3		12	
Unexposed	16		15		6	

logistic regression model gave no hint of this lack of positivity and provided estimates which smoothed over these empty cells (6). The authors of both papers are to be commended for examining this often-ignored assumption in their analyses.

Cheng et al. note that for some levels of the confounders, the exposure ''is empirically deterministic. . .so it makes little sense to determine the effect of [exposure] within those groups'' (5, p. **656**). We agree with the authors that if exposure is determined for some levels of the confounders, inference may be ill-advised because of nonpositivity. However, as they describe, for some levels of the confounders the probability of the *outcome* (not the exposure) becomes a certainty (see Figures 1 and 2 in Cheng et al. (5)). Admirably, the authors present the risk ratio as a function of the propensity score, showing effect measure modification graphically in their Appendix Figure. While important, a lack of variability in the outcome is not a concern of positivity.

A simple example, presented in Table 2, highlights the distinction between nonpositivity and the challenge described by Cheng et al. (5). In section A of Table 2, we show data in which the crude risk difference for an exposure X and an outcome Y is confounded by Z, and there are participants at both levels of X within both strata of Z: there is positivity. In section B of Table 2, we show a situation similar to that described by Cheng et al., in which there are participants at both levels of the exposure X within both strata of Z (again, positivity) but participants at only 1 level of the outcome Y within 1 stratum of Z (i.e., a zero column when $Z = 0$). Thus, stratum-specific risk differences by Z in Table 2, section B, show a risk difference of 0.00 for $Z = 0$ and 0.05 for $Z = 1$; the effect of X on Y is nonuniform by strata of Z. However, both quantities are estimable. Lastly, in section C of Table 2, we show an example of nonpositivity. Here, there are participants at both levels of Y within all strata of Z, but when $Z = 0$ there are no persons with $X = 1$. The risk difference in section C is inestimable in the $Z = 0$ stratum because estimation requires division by 0. Therefore, the concern expressed in section B of the table is orthogonal (in a fashion) to the issue of positivity.

How concerned should the practicing epidemiologist be with positivity in epidemiologic analysis? It seems clear that deterministic violations of positivity are of serious concern;

Table 2. Examples of Data Illustrating A) Positivity With a Uniform Risk Difference, B) Positivity With a Risk Difference Across Confounder Strata, and C) Nonpositivity[a]

Panel			Stratum-Specific Data				Total Data	
A)								
		Z	0		1		All *Z*	
		Y	1	0	1	0	1	0
	X	1	20	180	160	640	180	820
		0	40	760	30	170	70	930
RD (Wald 95% CI)			0.05 (0.01, 0.09)		0.05 (–0.01, 0.11)		0.11 (0.08, 0.14)	
B)								
		Z	0		1		All *Z*	
		Y	1	0	1	0	1	0
	X	1	0	200	160	640	160	840
		0	0	800	30	170	30	970
RD (Wald 95% CI)			0.00 (–0.01, 0.01)[b]		0.05 (–0.01, 0.11)		0.13 (0.10, 0.16)	
C)								
		Z	0		1		All *Z*	
		Y	1	0	1	0	1	0
	X	1	0	0	160	640	160	640
		0	60	940	30	170	90	1110
RD (Wald 95% CI)			NA (NA)		0.05 (–0.01, 0.11)		0.125 (0.09, 0.16)	

Abbreviations: CI, confidence interval; NA, not applicable; RD, risk difference.

[a] In all 3 sections of the table, X is the exposure, Y is the outcome, and Z is a confounder. Asymptotic Wald 95% CIs were calculated with SAS 9.2 (SAS Institute Inc., Cary, North Carolina).

[b] 0.5 was added to both 0 cells to calculate 95% confidence intervals for a valid point estimate of 0.00.

as in the case of assessing the effect of hysterectomy in men, they can lead to nonsensical statements of effect. Practicing epidemiologists should strive to avoid such deterministic violations of positivity, primarily through careful statement of the question at hand.

Random violations of positivity are more subtle and more difficult to assess. Messer et al. (6) discovered nonpositivity using a tabular approach in their analysis. When it is feasible, such a tabular approach quickly reveals empty cells. Continuous data can make such a tabular approach impractical, however. One alternative is to examine continuous variables for positivity by categories or quantiles (7) and proceed as with categorical variables. Such an approach was undertaken by Messer et al. (6), as well as by Cole and Hernán (2). However, choices about how to categorize a continuous variable can create or dispel random violations of positivity, without changing the underlying data, as Messer et al. point out in their discussion (6). In a population with large numbers of exposed participants, we might feel confident that modeling age in 10-year categories would result in exposed and unexposed persons in every category; but if we modeled age more finely (in 5-year categories, for example), we are more likely to find values of age at which positivity is violated. In truly continuous data, random violations of positivity become a certainty. In Table 1, for example, the observed nonpositivity (section A of the table) is eliminated when our age categories are coarsened to 31–40, 41–50, and 51–60 years (section B). More complicated still is the situation in which positivity is violated only in certain combinations of 2, 3, or even more confounders. In such settings, tabular analysis may be intractable even without continuous confounders.

Some have argued that positivity should be assessed only after a final model has been decided upon, so that important confounders are not ignored (3). We prefer to consider the tradeoff between bias due to positivity violations and bias due to confounding when deciding upon a final model (2). For instance, one may wish to incur a small amount of confounding bias to ensure against a large amount of bias due to nonpositivity, or vice versa. This is illustrated in Table 1. The categorization of age in Table 1, section A, would finely control for age but exhibits nonpositivity. Conversely, the coarser categorization of age in Table 1, section B, may result in increased residual confounding by age but does not suffer from nonpositivity. Methodological approaches are needed in order to weigh the resultant biases and make such decisions.

Once detected, the epidemiologist can deal with violations of positivity in several ways. The simplest solution is restriction. While easy and effective, restriction has the effect of altering the target population for inference. This method is the one implicitly favored by epidemiologists using propensity scores, who match or ‘‘trim’’ their data to avoid regions of propensity score nonoverlap. Another approach, one that is most appropriate for random violations of positivity that are surrounded by regions in which positivity holds, is to interpolate to areas of nonpositivity.

One possible approach to dealing with questions of positivity is as follows. We first ask whether there are covariates $\mathbf{V}$ for which $\Pr(X = x|\ \mathbf{V}) \approx 0$. If not, there are no potential problems with positivity. If so, then for variables $\mathbf{V}$ that are not confounders, one should look to the literature on exposure opportunity (8–10). For confounders $\mathbf{Z} \subseteq \mathbf{V}$ that are not time-varying, the best method for dealing with nonpositivity depends on the circumstances. If the nonpositivity is both random and internal (e.g., positivity at ages 36–40 and 46–50 years but not at ages 41–45 years), cautious interpolation or smoothing over the region of nonpositivity is reasonable. In such cases, restriction may prove more difficult, not least due to clearly defining the altered estimand. If the nonpositivity is random and external (e.g., no positivity under age 36 years), extrapolation is possible but often ill-advised. In such cases, restricting inference to persons aged 36 years or more may be a prudent approach. If the nonpositivity is deterministic, however, restriction can be recommended as an appropriate approach in many cases. Lastly, nonpositivity by a time-varying confounder poses an analytic challenge. In such cases, *g*-estimation of a structural nested model (11) or *g*-computation (12) may be a way forward, but more research is needed.

In conclusion, issues of positivity are as old as observational data analysis but have been formalized only relatively recently. Hopefully, recent formalization will lead toward better understanding of and accounting for positivity in epidemiologic research.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Daniel Westreich, Stephen R. Cole).

Dr. Daniel Westreich was supported by National Institutes of Health/National Institute of Allergy and Infectious Diseases grant T32-AI-07001. Dr. Stephen R. Cole was supported in part by grants R01-AA-01759 and P30-AI-50410.

The authors thank Elizabeth Stuart and Constantine Frangakis for expert advice.

Conflict of interest: none declared.

REFERENCES

1. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006; 60(7):578–586.
2. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008; 168(6):656–664.
3. Mortimer KM, Neugebauer R, van der Laan M, et al. An application of model-fitting procedures for marginal structural models. *Am J Epidemiol*. 2005;162(4):382–388.
4. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009; 20(1):3–5.
5. Cheng YW, Hubbard A, Caughey AB, et al. The association between persistent fetal occiput posterior position and perinatal outcomes: an example of propensity score and covariate distance matching. *Am J Epidemiol*. 2010;171(6): 656–663.

6. Messer LC, Oakes JM, Mason S. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *Am J Epidemiol*. 2010;171(6):664–673.
7. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24(2):295–313.
8. Poole C. Critical appraisal of the exposure-potential restriction rule. *Am J Epidemiol*. 1987;125(2):179–183.
9. Poole C. Exposure opportunity in case-control studies. *Am J Epidemiol*. 1986;123(2):352–358.
10. Schlesselman JJ, Stadel BV. Exposure opportunity in epidemiologic studies. *Am J Epidemiol*. 1987;125(2):174–178.
11. Robins JM. Structural nested failure time models. In: Armitage P, Colton T, eds. *The Encyclopedia of Biostatistics*. Chichester, United Kingdom: John Wiley & Sons Ltd; 1997: 4372–4389.
12. Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chronic Dis*. 1987;40(suppl 2): 139s–161s.