Journal of Clinical Epidemiology 139 (2021) 264–268

Journal of Clinical Epidemiology

Noncollapsibility, confounding, and sparse-data bias. Part 2: What should researchers make of persistent controversies about the odds ratio?

Sander Greenland

Department of Epidemiology and Department of Statistics, University of California, CA, Los Angeles

Received 14 March 2021; Received in revised form 4 June 2021; Accepted 4 June 2021; Available online 11 June 2021

Abstract

A previous note illustrated how the odds of an outcome have an undesirable property for risk summarization and communication: Noncollapsibility, defined as a failure of a group measure to represent a simple average of the measure over individuals or subgroups. The present sequel discusses how odds ratios amplify odds noncollapsibility and provides a basic numeric illustration of how noncollapsibility differs from confounding of effects (with which it is often confused). It also draws a connection of noncollapsibility to sparse-data bias in logistic, log-linear, and proportional-hazards regression.

Keywords: Causality; Collapsibility; Confounding; Logistic regression; Noncollapsibility; Odds Ratio; Rate Ratio; Simpson's paradox; Sparse-data bias

1. Noncollapsibility vs. confounding

Part 1 [13] explained how noncollapsibility is a noncausal phenomenon in which a measurement on a group does not equal a designated average of the same measurement over its constituents, as illustrated by how group odds are not simple averages of individual odds when the odds vary across individuals. Let us now move on to an even more odd phenomenon: The odds ratio measuring a treatment effect on a group need not equal *any* average of its subgroup odds ratios, *even if there is no confounding*; in particular, we can have noncollapsibility over subgroups defined by a risk factor even if that factor is not a confounder ([26], p. 580). For simplicity I will assume the treatment increases outcome frequency, but by interchanging the labels "treated" and "untreated" the example can be seen to also apply to preventive treatments.

Confounding of effects by a covariate has been defined in many ways. Much of the literature before the present century defined it loosely as confusing or mixing effects of baseline (pretreatment) covariate effects with treatment effects. For that effect mixing to occur, the baseline covariate must have effects of its own on the patient outcome (effects not mediated by the treatment) and must be associated with the treatment at baseline (the time of treatment). This loose definition will be used to illustrate the distinction between confounding and noncollapsibility with a simple example (there are many variations and subtleties in more precise definitions, for example, see [8,25] Ch. 6, [17]).

Consider again the example in part 1 of a study of 150 patients over a fixed time period following a treatment with high mortality (such as an experimental intervention for intensive-care patients), with 45 deaths among 50 males and 30 deaths among 100 females. Suppose further that each patient was matched to an untreated control patient who was so similar that this match was prognostically indistinguishable from (or *exchangeable* with) the treated patient, being of the same birth year, sex, BMI, etc. This would result in 50 males and 100 females in the control group, just as in the treated group. This balance (equality) of the sex ratio between the treated and control groups means that any difference in outcome between the treated and control group cannot be explained (or confounded) by a difference in sex ratio.

If there were 30 male and 10 female control deaths, the numbers would be as in Table 2. A key feature of this table is that the collapsed odds ratio is closer to 1 than both the male and female odds ratios, and so cannot be any kind of average of the latter two. This qualitative difference between what is seen conditional on sex (within the subgroups) and marginally (unconditionally) in the total cannot be ascribed to a sex imbalance between the

Conflict of Interest None.

E-mail address: lesdomes@ucla.edu.

https://doi.org/10.1016/j.jclinepi.2021.06.004

Table 1. Numeric example of odds-ratio noncollapsibility without confounding

	Males		Females		Collapsed	
	Treated	Untreated	Treated	Untreated	Treated	Untreated
Died	45	30	30	10	75	40
Survived	5	20	70	90	75	110
Totals	50	50	100	100	150	150
Proportions (risks)	0.90	0.60	0.30	0.10	0.500	0.267
Odds	9/1	3/2	3/7	1/9	1/1	4/11
Odds ratios	6.0		3.9		2.75	
Risk ratios	1.5		3.0		1.875	
Risk differences	0.30		0.20		0.233	

treated and untreated, because there is none. It thus cannot represent confounding by sex because there is none, even though sex is a strong risk factor for death among these patients and is not affected by the treatment. This leads to the apparent paradox that the treatment more than triples the odds of death in both the male and female subgroups, yet falls short of tripling the odds of death when considering the group as a whole.

This oddity of noncollapsibility without confounding has been the subject of some 40 years of discussion; see Didelez and Stensrud [4] and Daniel et al. [3] for recent reviews. To see the problem, note that odds of 1 in the total (collapsed) treated group are an average of the male and female odds weighted by 5 and 70,

$$\{5(9/1) + 70(3/7)\}/(5 + 70) = (45 + 30)/75 = 1$$

which gives only 5/75 = 6.7% of the weight to males. In contrast, the odds of 4/11 in the total untreated group weights the male and female odds by 20 and 90,

$$\{20(3/2) + 90(1/9)\}/(20 + 90) = (30 + 10)/110 = 4/11 = 0.36$$

which gives males by nearly 3 times as much weight, 20/110 = 18.2%. Because the males have a higher odds ratio but a much lower proportion of survivors than do the females, their lower weight in the total treated odds compared to the untreated odds results in the collapsed odds ratio being less than both the male and female odds ratios. This noncollapsibility is not however confounding because it is produced by an effect of treatment on the weights, rather than by a pre-treatment (baseline) difference between the treatment groups.[1]

Hazard (incidence) rate differences and ratios can also suffer from noncollapsibility without confounding, albeit not as severely as the odds ratio [7,8,15,23,29]. Such odd behavior of odds-based and hazard-based measures is exactly what one should expect whenever the treatment and covariates affect survival [7], including in randomized trials [15].[2] In contrast, risk differences and risk ratios are collapsible if there is no confounding: As illustrated in Table 1, the collapsed risk difference is an average of the covariate-specific risk differences weighted by subgroup sizes,

$$\{100(0.30) + 200(0.20)\}/(100 + 200) = (30 + 40)/300 = 0.233$$

while the collapsed risk ratio is an average of the covariate-specific risk ratios weighted by the number of deaths in each untreated subgroup,

$$\{30(1.5) + 10(3.0)\}/(30 + 10) = (45 + 30)/40 = 1.875.$$

The confusion of noncollapsibility and confounding has been partly encouraged by the identification of both with "Simpson's paradox", although the original version of the paradox was closer to noncollapsibility in form. Most important is that noncollapsibility without confounding can arise for *any* measure of association (not just odds ratios) when the covariate is affected by the treatment. See [8,11,16,25–27] for further discussion of the distinctions among noncollapsibility, confounding, and "Simpson's paradox".

2. Uncommon outcomes can become common in sparse data

As with confounding, noncollapsibility can afflict arbitrarily large data sets (as illustrated by multiplying all the numbers in Table 2 by any large number and noting that the odds ratios do not change). The extent of noncollapsibility is also direct function of the variation in the outcome proportions across both the treatment and the subgrouping. In particular, if, *in both treatment arms and across all*

[1] This noncollapsibility can be circumvented if instead we adopt the same weighting system for the treated and untreated odds; the resulting ratio is then an average of the stratum (subgroup) specific odds ratios weighted by the odds in the untreated times the weights used for the odds. Extensions of this idea lead to formulas for separating noncollapsibility (nonlinearity) and confounding effects on the odds ratio [19,24].

[2] Odds ratios and rate ratios can also suffer from confounding by a covariate and yet be collapsible over that covariate, although examples of that are somewhat contrived, involving perfect cancelation of noncollapsibility and confounding effects [24].

subgroups, those proportions are always "small" (eg, under 10%) or else always "large" (eg, over 90%), the extent of noncollapsibility will be limited. That can be illustrated by leaving the number of deaths in Table 2 unchanged but multiplying the total number of patients in each subgroup by 10. Then the proportions are all cut to a tenth of those shown in Table 2, to less than 0.10; and while the risk ratios remain unchanged, the odds ratios are now far smaller (within 10% of the risk ratios[3]) with male and female odds ratios of 1.5 and 3.1 and a collapsed odds ratio of 1.9 falling between them. Parallel results apply to hazard (incidence) rate ratios.

Unfortunately, large variations in the outcome proportions can arise when we increase division into subgroups defined by covariates that further predict the outcome (as is often done to improve confounding control). These finer divisions even lead to subgroups in which the outcome is either absent or present for everyone in a given treatment arm and thus the outcome variation is maximal [13]. As a result, odds-ratio noncollapsibility can increase dramatically as we stratify ever more finely on predictors of the outcome. This problem will also manifest itself in odds ratios derived from logistic and log-linear models for count data, in conditional logistic models for matched data, and in hazard rate ratios from proportional-hazards models for survival data [9].

Through this mechanism, the fine covariate subgrouping imposed by modern multiple-regression models may aggravate noncollapsibility. For example, the effective number of subgroups underlying a multiple-logistic model with 10 binary covariates (not an unusual number) is 2^{10} or over a million. With fewer than a million patients, one is likely to find many subgroups with only a few patients; with only a few thousand patients most subgroups will be small, with odds ratios either 0 or undefined due to zero cells. This phenomenon is moderated by the smoothing effect (shrinkage) produced by model constraints (such as additivity and response linearity on the logit scale) which pulls the estimated subgroup odds ratios to finite values. Nonetheless, noncollapsibility increases with the ability of model covariates to predict the outcome. The typical result is movement of estimated odds ratios away from 1 as more predictors are added to the model, even when no further confounding or other bias is being removed.

This estimate inflation from "too many covariates chasing too few data points" is often discussed under the topic of *sparse-data bias*, and can be quite large in real-data examples [9,12].[4] Yet it usually goes unnoticed or else is misattributed to confounding removal and thus treated as if it were bias removal, when it is instead an artefact of overadjustment. While it can be mitigated somewhat through use of special methods (such as conditional or partial likelihood), these methods can themselves suffer from estimate inflation when they attempt to adjust for too many covariates; even randomized trials that attempt to adjust for too many covariates relative to their size are not exempt. If nonetheless one desires to adjust for many covariates, more elaborate methods (such as careful use of penalization, propensity scoring, or some combination) will be needed to prevent the bias.

3. Discussion

Noncollapsibility is a purely numeric averaging failure that can be found in basic measures of occurrence or frequency as well as in comparisons, whereas confounding concerns causal comparisons and requires a causal framework to define and illustrate properly. Writings lacking that framework are still prone to confuse the two phenomena, which can be especially harmful when increased covariate adjustment leads to estimate inflation and that is mistaken for confounding control. These considerations lead to one answer to the question in the title of this paper: Odds ratios can be misleading if used to select covariates for confounding adjustment, and (except when they approximate risk ratios) they should not be treated as the final target of a statistical analysis or as average or typical effects.

What then should be done if the outcome is so common that odds ratios no longer approximate risk ratios? Unfortunately, the most popular option uses a formula that purports to convert odds ratios directly to risk ratios [33] produces biased risk ratios when fed ordinary logistic model odds ratios [21]. A less biased option is to use a model whose coefficients are interpretable as log risk ratios [30]; unfortunately these models tend to be highly unrealistic when the outcome is common (because they do not restrict the range of the fitted proportions to between 0 and 1). Thus, my preferred solution is to stay with the logistic model but then use the fitted risks from the model to construct collapsible measures, such as covariate-specific and weighted-average (standardized) risk differences and risk ratios [10,14,20].

In summary, noncollapsible measures change upon adjustment for outcome predictors even if the latter aren't confounders or effect modifiers, adding to problems

[3] This approximation of odds ratios to proportion ratios was first noted by Cornfield [2].

[4] Strictly speaking, *sparse-data bias* refers to the tendency of the estimates to concentrate away from the true value of the parameter on repeated sampling as a result of fitting a model with too many free parameters relative to the sample size. For example, in a pair-matched trial there can be no confounding by the matching factors. Nonetheless, the N pairs form N subgroups in which the sample odds ratios are undefined. To deal with this problem, the underlying subgroup odds ratios are usually assumed to be constant. But with no further constraint the resulting sampling model has N+1 free parameters, and fitting by unconditional maximum likelihood results in an estimator that converges to the square of the correct odds ratio as the N increases (Anderson [1], p. 69). In this case, switching to conditional maximum likelihood can restore convergence to the correct odds ratio, but more generally it too can suffer from sparse-data bias [9].

of transportability, noncomparability across studies, and sparse-data bias (which can arise even for "rare" outcomes). There are of course many other considerations that are crucial for proper adjustment and interpretation of effect estimates, especially the causal ordering of variables and the effects targeted for estimation [5,6,17,25,31,32]. Nonetheless, it is hoped the present numeric introduction to noncollapsibility will help readers understand some of the more subtle distinctions and reasons for preferring measures comparing proportions (risks) rather than odds whenever the odds do not approximate risks (eg, [3,4,8,11,16–19,22,24,28,34]).

4. Further reading

Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. Am J Epidemiol 2004;160:301-305. *Explains the principles behind correct conversion of outputs from risk-regression models into estimates of average effects (valid effect standardization).*

Greenland S, Mansournia MA, Altman DG. Sparse-data bias: A problem hiding in plain sight. Br Med J 2016;353:i1981, 1-6. *Describes the bias produced in logistic-regression estimates when there are too many adjustment covariates in the model relative to the amount of data used to fit the model.*

Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. Int J Epidemiol 2011;40:780– 785. *Describes the connection of "Simpson's Paradox" to modern notions of noncollapsibility and confounding and how the paradox is resolved by adding causal structure to the statistical relations.*

Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Statist Meth Med Res 2016;25(5):1925-1937. *Explains how to quantify the relative contributions of noncollapsibility and confounding to changes upon covariate adjustment.*

Westreich D, Greenland S. The table-2 fallacy: Presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013:177;292-298. *Explains why confounder coefficients in regression models often do not measure the effects of the confounder on the outcome (eg, when the exposure is an intermediate between the confounder and the outcome, or the confounder is itself confounded even if the exposure is not).*

References

[1] Andersen EB. Conditional inference and models for measuring. Copenhagen: Mentalhygienisk Forlag; 1973. p. 69.

[2] Cornfield J. A method of estimating comparative rates from clinical data: application to cancer of the lung, breast and cervix. J Natl Cancer Inst 1951;11:1269–75.

[3] Daniel R, Zhang J, Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biomet J 2021;63:528–57.

[4] Didelez V, Stensrud M. On the logic of collapsibility for causal effect measures. Biometrical J 2021 in press.

[5] Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd ed.. Philadelphia, Lippincott-Wolters-Kluwer; 2008. p. 183–209.

[6] Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. Int J Epidemiol 1980;9:361–7.

[7] Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 1996;7(5):498–501.

[8] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Statist Sci 1999;14(1):29–46.

[9] Greenland S, Schwartzbaum JA, Finkle WD. Problems from small samples and sparse data in conditional logistic regression analysis. Am J Epidemiol 2000;151:531–9.

[10] Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case–control studies. Am J Epidemiol 2004;160:301–5.

[11] Greenland S, Pearl J. Adjustments and their consequences - Collapsibility analysis using graphical models. Int Statist Rev 2011;79(3):401–26.

[12] Greenland S, Mansournia MA, Altman DG. Sparse-data bias: A problem hiding in plain sight. Br Med J 2016;353(i1981):1–6.

[13] Greenland S. Noncollapsibility, confounding, and sparse-data bias. Part 1: The oddities of odds. J Clin Epidemiol 2021; in press.

[14] Greenland S. Introduction to Regression Modeling. Ch. 21. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd ed. Philadelphia, Lippincott-Wolters-Kluwer; 2008. p. 418–55.

[15] Hernán MA. The hazards of hazard ratios. Epidemiology 2010;21(1):13–15.

[16] Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. Int J Epidemiol 2011;40:780–5.

[17] Hernán MA, Robins JM. Causal inference: What If. Boca Raton, FL: Chapman & Hall/CRC; 2020. p. 54.

[18] Huitfeldt A, Stensrud MJ, Suzuki E. On the collapsibility of measures of effect in the counterfactual causal framework. Emerg Themes Epidemiol 2019;16(1):1.

[19] Janes H, Dominici F, Zeger S. On quantifying the magnitude of confounding. Biostatistics 2010;11:572–82.

[20] Joffe MM, Greenland S. Estimation of standardized parameters from categorical regression models. Stat Med 1995;14:2131–41.

[21] Karp I. Re: "Estimating the relative risk in cohort studies and clinical trials of common outcomes" (letter). Am J Epidemiol 2014;179(8):1034–5 179.

[22] Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. Epidemiology 2015;26(4):466–72.

[23] Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Analysis 2013;19:279–96.

[24] Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Statist Meth Med Res 2016;25(5):1925–37.

[25] Pearl J. Simpson's paradox, confounding, and collapsibility. Ch. 6 in Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge, MA: Cambridge University Press; 2009. p. 173–200.

[26] Samuels ML. Matching and design efficiency in epidemiological studies. Biometrika 1981;68:577–88.

[27] Samuels ML. Simpson's paradox and related phenomena. J Am Statist Assoc 1993;88:81–8.

[28] Shrier I, Pang M. Confounding, effect modification and the odds ratio: Common misinterpretations. J Clin Epidemiol 2015;68(5):470–4.

[29] Sjölander A, Dahlqwist E, Zetterqvist J. A note on the noncollapsibility of rate differences and rate ratios. Epidemiology 2016;27(3):356–9.

[30] Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. Am J Epidemiol 2005;162(3):199–200.

[31] VanderWeele TJ. Explanation in Causal Inference. New York: Oxford; 2015.

[32] Westreich D, Greenland S. The table-2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013;177:292–8.

[33] Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 1998;280(19):1690–1 Nov 18.

[34] Xiao M, Chu H, Cole SR, Chen Y, MacLehose RF, Richardson DB, Greenland S. Odds ratios are far from "portable" - A call to use realistic models for effect variation in meta-analysis. J. Clin. Epidemiol 2021 in press.