ELSEVIER

Journal of Clinical Epidemiology 138 (2021) 178–181

Journal of Clinical Epidemiology

KEY CONCEPTS SERIES

Noncollapsibility, confounding, and sparse-data bias. Part 1: The oddities of odds

Sander Greenland

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

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

Abstract

To prevent statistical misinterpretations, it has long been advised to focus on estimation instead of statistical testing. This sound advice brings with it the need to choose the outcome and effect measures on which to focus. Measures based on odds or their logarithms have often been promoted due to their pleasing statistical properties, but have an undesirable property for risk summarization and communication: Noncollapsibility, defined as a failure of the measure when taken on a group to equal a simple average of the measure when taken on the group's members or subgroups. The present note illustrates this problem with a basic numeric example involving the odds, which is not collapsible when the odds vary across individuals and are not low in all subgroups. Its sequel will illustrate how this problem is amplified in odds ratios and logistic regression.

Keywords: Causality; Collapsibility; Confounding; Noncollapsibility; Odds ratio; Rate ratio; Simpson's paradox

1. Introduction

One of the oldest preventives against misinterpretations of statistics is to switch focus from testing to estimation of effect sizes [1–5]. Alas, many journals perpetuate old significance-testing fallacies by having authors conclude "no significant effect observed" because the 95% interval estimate contained the null, or by allowing them to claim a "significant effect" was found because the interval excluded the null. Doing better requires looking at the full range of the interval with a sound idea of what is clinically important. But there are many ways to measure effect sizes and much controversy about which measure to use. Unfortunately, some choices can be opaque and even misleading for clinical practice. For example, so- called "standardized" effect measures and correlation coefficients confound clinically irrelevant study features with actual effects on patients [6,7].

How then should we measure effects? While some issues have been framed in terms choosing difference ("absolute") vs. ratio ("relative") measures, others can be seen arising from the more basic choice of how to quantify outcomes of treatment before comparing them. I will here focus on just one consideration, noncollapsibility, which has weighed against measures based on odds (and to a lesser extent measures based on hazard rates) when the outcome under study is an event common within some subgroups. The literature on the topic is now vast, delving into the many subtleties, hence I will not attempt to provide a history or theoretical review.

Instead, I will illustrate the numeric properties of odds that produce noncollapsibility, while part 2 [8] will illustrate how those properties lead to more severe noncollapsibility of odds ratios. Noncollapsibility can render odds ratios problematic for estimating effects in special subgroups or patients, and can be amplified when adjusting for many covariates; it also raises the question of how to present results from logistic regression, whose exponentiated coefficients represent odds ratios. It is hoped that this pair of papers will aid readers in approaching more detailed analyses of the distinction and the reasons for preferring measures comparing proportions (risks) rather than odds whenever the odds do not approximate the risks (eg, [9,10] Ch. 6, [11] p. 62, [12–20] p. 54, [21,22]). Such measures can be computed from logistic-regression output as well as from other types of models ([23,24] p. 442-446).

2. An oddity of odds: Undefined individual outcomes and no simple averaging

Noncollapsibility can refer to either estimates computed from a single data set or the corresponding parameters in the data-generating distribution. To minimize abstractions I

Conflict of interest: None.

E-mail address: lesdomes@ucla.edu.

https://doi.org/10.1016/j.jclinepi.2021.06.007
0895-4356/© 2021 Elsevier Inc. All rights reserved.

will focus on observed frequencies of the simple outcome of death vs. survival in the year following a one-time treatment with high mortality (such as an experimental transplant or late-stage oncotherapy). For a single patient this outcome is easily coded as 1 for death and 0 for survival, called a binary *death indicator*; reversing the code to 0 for death, 1 for survival gives us a binary *survival indicator*.

Suppose we have a study group of 150 treated patients, of whom 75 die within a fixed study period. One summary for the group is the proportion dying (or observed mortality risk), 75/150 = 0.50, which equals the simple arithmetic mean of the death indicators for all 150 patients. Another would be the proportion surviving, 75/150 = 0.50, which equals the arithmetic mean of the survival indicators. These have familiar, transparent interpretations of 50% dying, 50% surviving. Another choice would be the observed odds of death (or mortality odds) 75/75 =0.50/0.50 = 1.

The general formula for the odds of an outcome that occurs in a proportion p of patients is p/(1−p). At first it might seem the choice between proportions and odds is unimportant: We can not only get odds from proportions, but we can also get back from an odds to the proportion via the formula odds/(1+odds) = p, which in our example gives 1/(1+1) = 0.50. If however we want to get back to individual patient effects, there is a devil in the details: We know the proportion p is a simple average of individual outcome indicators. For what individual outcome does the odds provide a simple average? For real data composed of outcome indicators, the answer is: None.

To see why, imagine each individual patient forming their own subgroup of size 1, defined by their genetic code and entire life history. The proportion of this one-patient subgroup dying in the month after treatment is either 0 (if the patient survives) or 1 (if they die). The one-patient subgroup odds is then either 0/1 = 0, meaning that death was as uncommon as can be (ie, no one died in this subgroup) or 1/0, meaning that death is as common as can be (ie, no one survived in this subgroup). Now 1/0 is either undefined (as in most math and computer systems) or represents infinite odds (no chance of escaping the outcome). It follows that if any patient dies, the average of these individual odds across the original group will itself be undefined or infinite, which does not equal or even hint at the observed group odds of 1.

One attempt to fix this lack of relation of group odds to individual odds is called *coarsening*: Subdivide no further than subgroups that are large enough so that there both deaths and survivals within them all, making it possible to compute odds within each subgroup that are neither infinite nor zero. Alas, this maneuver does not make the group odds a simple average of individual odds. To see this, suppose our original group of 150 patients was composed of 50 males and 100 females, and that 45 of the deaths were male and 30 were female. The numbers are summarized in Table 1.

Table 1. Numeric example of odds noncollapsibility

	Male	Female	Total
Died	45	30	75
Survived	5	70	75
Total	50	100	150
Proportion (risk)	0.90	0.30	0.50
Odds	9/1 = 9.00	3/7 = 0.43	1/1 = 1

The proportion males and females dying are 0.90 and 0.30, each of which is a simple average of the death indicators in its group:

$$\{45(1) + 5(0)\}/50 = 45/50 = 0.90 \quad \text{and}$$
$$\{30(1) + 70(0)\} = 30/100 = 0.30.$$

Furthermore, the proportion dying in the undivided (total) group is a simple average of the death indicators across the groups:

$$\{45(1) + 5(0) + 30(1) + 70(0)\}$$
$$= \{75(1) + 75(0)\}/100 = 75/100 = 0.50.$$

This equality reflects how proportions are *collapsible* for simple averages of individual outcomes across subgroups: The proportion for the total is the same whether or not we separate (stratify) the outcomes by subgroup when averaging the individual outcomes; that is, the average proportion does not change if we ignore the covariate, even though the covariate strongly predicts the outcome.

In contrast, the odds of dying among males and females are 45/5 = 9/1 = 9 and 30/70 = 3/7 =0.43. As we saw above, these cannot be averages of individual odds because the latter are either 1/0 which is undefined, or 0/1 = 0. Suppose to overcome this deficiency we assign each individual the odds for the narrowest subgroup in which they fall (as might be obtained from a regression of outcome on sex). Then every male would be assigned an odds of 9, every female an odds of 3/7, and the simple average individual odds for the total would be

$$\{50(9) + 100(3/7)\}/150 = 3.3$$

which is notably larger than the odds of 1 for the total group computed without stratifying on sex. This inequality reflects how, unlike the proportion, odds are *noncollapsible* for simple averages over individuals if the odds vary across individuals.

Several central points can be seen in examples like that above:

(1) The concept of noncollapsibility is not a causal one, but rather a much more basic and general arithmetical one: There is no causal effect being measured in the above example, yet (unlike for proportions) we can see that the odds for the total group fails to equal the average individual odds; this would be so even if the odds variation was not caused by the

subgrouping variable. Thus noncollapsibility is not a bias unless we mistakenly take the odds for the total group as showing the simple average odds.

(2) Noncollapsibility can afflict arbitrarily large data sets; it is not just a problem of small numbers. For example, if we multiplied the counts in Table 1 by 10, we would see exactly the same proportions, odds, and degree of noncollapsibility.

(3) Noncollapsibility is directly proportional to the sizes of the odds. Very low odds will approximate proportions (because for very small p, $1-p \approx 1$ and so $p/(1-p) \approx p$), and thus the odds will be approximately collapsible for their simple mean when everyone has low odds. This can be illustrated by leaving the number of deaths in Table 1 unchanged but multiplying the totals by 10: The proportions then become 0.090 and 0.030, which are only 10% and 3% less than the odds of 0.099 and 0.031.[1] Conversely, however, if the odds are high in any subgroup they will diverge from the proportions and noncollapsibility of the odds can ensue.

(4) If we continue to subdivide the data ever more finely on covariates that further predict the outcome, we will see subgroups with ever higher odds and thus see ever more noncollapsibility, eventually reaching subgroups that have only 1 member and thus 0 or undefined (or infinite) odds – the most extreme sparse data.

Following up the last point, statistical modeling typically introduces more covariates but replaces the observed outcome indicator for each patient with a hypothetical individual probability p derived from a risk model. Provided the model keeps p strictly between 0 and 1 (as in logistic models) it will supply a finite, nonzero individual-patient odds from the formula $p/(1-p)$. As seen above, however, this device does not make the total-group odds equal to a simple average of the individual odds [25].

Noncollapsibility also afflicts other common frequency measures, albeit not as severely as for odds. In the above example the natural logs of the odds (or logit) for males and females are $\ln(9) = 2.20$ and $\ln(3/7) = -0.85$ with an individual average of 0.17, exceeding the logit for the total, $\ln(1) = 0$. The cumulative hazards $-\ln(1-p)$ for males and females are $-\ln(1-0.90) = 2.30$ and $-\ln(1-0.30) = 0.36$ with an individual average of 1.01, exceeding the value for the total, $-\ln(1-0.50) = 0.69$. Finally, the above example is easily modified to show that person-time incidence or mortality rates (cases/person-time) and their logs can suffer similar noncollapsibility when (as usual in clinical trials) the person-time distribution is affected by survival [26]: relabel the "Survivor" row as "Person-years", leave the "Total" row as the size of the starting population, and relabel the "Odds" row as "Rate" with units of inverse person-years.

The concept of noncollapsibility can be extended to weighted averages ([10], Ch. 6) and regression coefficients ([9], sec. 4.2). For example, the total-group mortality proportion is an average of subgroup proportions weighted by the total subgroup numbers:

$$\{50(0.90) + 100(0.30)\}/(50+100) = 0.50$$

Similarly, the total-group odds is an average of subgroup odds weighted by the number of subgroup survivors:

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

illustrating that the odds is indeed collapsible for this particular weighted average; in a parallel fashion, the total-group mortality rate is an average of subgroup rates weighted by subgroup person-time [25,26]. Unfortunately, weighted averaging by survivors or person-time has some unsettling consequences for odds ratios from logistic regression and for rate ratios from survival regressions, even in randomized trials: If the treatment has an effect, it will change the numbers surviving and the person-time, and thus create a difference in weighting between the untreated and treated groups. Part 2 will discuss these problems and the relation of the resulting noncollapsibility to notions of confounding, regression adjustment, and "Simpson's paradox" [8].

Further reading

Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Statist Sci 1999;14(1):29–46. Gives some history behind concepts of confounding and collapsibility along with explanations of how they differ from one another and their relation to "Simpson's paradox."

Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology 1991;2:387-392. Explains why multiplying regression coefficients their covariate's standard deviation to produce "standardized coefficients" distorts comparisons among effect sizes.

Hernán MA, Robins JM. Causal inference: What If. Boca Raton, FL: Chapman & Hall/CRC, 2020. Provides a comprehensive and rigorous introduction to the burgeoning field of causal modeling for study design and data analysis.

Pearl J. Simpson's paradox, confounding, and collapsibility. Ch. 6 in Pearl J. Causality: Models, Reasoning, and Inference. Cambridge, MA: Cambridge University Press, 2000, 173-200. Provides an overview of the relations and distinctions among "Simpson's paradox," confounding, and collapsibility within the framework of formal causality theory.

[1] Their percent difference is most easily expressed using the odds: The ratio of the odds to the proportion is odds/p = $1/(1-p)$, hence the odds is $100(1/(1-p) - 1) = 100p/(1-p) = 100$•odds percent above the proportion.

Shrier I, Pang M. Confounding, effect modification and the odds ratio: Common misinterpretations. J Clin Epidemiol 2015;68(5):470-474. Explains why logistic and proportional-hazards (Cox) regression are unsuitable for the study of variation in causal effects (effect modification) due to their noncollapsible outputs.

References

[1] Yates F. The influence of statistical methods for research workers on the development of the science of statistics. J Am Stat Assoc 1951;46:19–34.

[2] Rothman KJ. A show of confidence. N Engl J Med 1978;299:1362–3.

[3] Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence. 2nd ed. London: BMJ Books; 2000.

[4] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337–50.

[5] Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p"<0.05. Am Stat 2019;73:1–19.

[6] Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. Am J Epidemiol 1986;123:203–8.

[7] Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology 1991;2:387–92.

[8] Greenland S. Noncollapsibility, confounding, and sparse-data bias. Part 2: what should researchers make of persistent controversies about the odds ratio? J Clin Epidemiol 2021.

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

[10] Pearl J. Simpson's paradox, confounding, and collapsibility. Ch 6 in Pearl J Causality: models, reasoning, and inference. Cambridge, MA: Cambridge University Press; 2000. p. 173–200.

[11] Ch. 4 Greenland S, Rothman KJ, Lash TL. Measures of effect and measures of association. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd ed. Philadelphia: Lippincott–Wolters- Kluwer; 2008. p. 51–70.

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

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

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

[15] Pang M, Kaufman JS. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Statist Meth Med Res 2013;25(5):1925–37.

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

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

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

[19] 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.

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

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

[22] 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.

[23] 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.

[24] Greenland S. Introduction to regression modeling. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology; 2008. p. 418–55. Ch. 21.

[25] Greenland S. Interpretation and choice of effect measures in epidemiologic analysis. Am J Epidemiol 1987;125:761–8.

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