American Journal of Epidemiology

Vol. 190, No. 2
DOI: 10.1093/aje/kwaa167
Advance Access publication:
September 5, 2020

Practice of Epidemiology

Do Case-Control Studies Always Estimate Odds Ratios?

Jeremy A. Labrecque*, Myriam M. G. Hunink, M. Arfan Ikram, and M. Kamran Ikram

* Correspondence to Dr. Jeremy A. Labrecque, Department of Epidemiology, Erasmus MC, PO Box 2040, 3000 CA, Rotterdam, The Netherlands (e-mail: j.labrecque@erasmusmc.nl).

Initially submitted November 12, 2019; accepted for publication August 11, 2020.

Case-control studies are an important part of the epidemiologic literature, yet confusion remains about how to interpret estimates from different case-control study designs. We demonstrate that not all case-control study designs estimate odds ratios. On the contrary, case-control studies in the literature often report odds ratios as their main parameter even when using designs that do not estimate odds ratios. Only studies using specific case-control designs should report odds ratios, whereas the case-cohort and incidence-density sampled case-control studies must report risk ratio and incidence rate ratios, respectively. This also applies to case-control studies conducted in open cohorts, which often estimate incidence rate ratios. We also demonstrate the misinterpretation of case-control study estimates in a small sample of highly cited case-control studies in general epidemiologic and medical journals. We therefore suggest that greater care be taken when considering which parameter is to be reported from a case-control study.

case-control studies; control sampling; incidence rate ratio; odds ratio; risk ratio

The case-control study is an important but often misunderstood study design. In our current understanding, a case-control study is nothing but an efficiently conducted cohort study achieved by sampling a subset of potential controls to get a measure of exposure distribution among them. Much has been written trying to clear the confusion in conducting and reporting on case-control studies (1–4), but there remain misunderstandings about how to interpret results from them. In 2008, Knol et al. (1) reviewed 150 case-control studies in 20 journals to survey which parameter they reported—odds ratios, risk ratios, or incidence rate ratios—and which parameter their design would have allowed them to estimate. They found that 90% of the studies reported only an odds ratio despite the fact that the majority used designs that estimate risk ratios or incidence rate ratios. It appears as though the attitude toward case-control studies is that it is always correct to present an odds ratio, but that with some designs the estimate also has a second interpretation either as a risk ratio or incidence rate ratio. Here, we aim to prevent further such confusion about the parameter estimated in case-control studies by explaining why many commonly used case-control designs do not, in fact, estimate odds ratios.

WHAT DOES A CASE-CONTROL STUDY ESTIMATE?

The findings of Knol et al. (1) should make epidemiologists pause. How is it that the parameter estimated in a case-control design—a design that is required learning to obtain a degree in epidemiology at any level—is misinterpreted in the majority of studies using it in the leading medical and epidemiologic journals? This odds ratio–centric view of case-control studies can also be seen in epidemiologic textbooks: "[I]n a case-control study the relative risk cannot be calculated directly" (5, p. 208) or, "the primary measure of effect in a case-control study is the odds ratio" (6, p. 45).

Much of this confusion is due to a mismatch between the statistical interpretation of a 2×2 table from a case-control study (Table 1) and its epidemiologic reality. Let us begin with the most often taught example, a case-control study conducted within a closed cohort where controls are sampled at the end of follow-up. To simplify things, let us imagine we have all the cases, meaning we know the true values in column $Y = 1$ of our 2×2 table. In this design, we sample a fraction, f, of the participants from column $Y = 0$. Therefore, we know the value of a and c as well as the values $b \times f$ and $d \times f$. If we know the sampling fraction (f), we

Table 1. Two-by-Two Table From a Fully Enumerated Cohort

Exposure	$Y = 1$	$Y = 0$	Total	Person-Time
$E = 1$	a	b	$a + b$	PT_1
$E = 0$	c	d	$c + d$	PT_0

Abbreviations: E, exposure; PT_0, person-time unexposed; PT_1, person-time exposed; Y, outcome.

can calculate b and d allowing us to calculate any measure of association. (7)

If we do not know f, what can we estimate knowing only $a, c, b \times f$ and $d \times f$? Most will answer, correctly, that we can estimate an odds ratio. This is where the mismatch between statistical interpretation and epidemiologic reality begins. When an estimate is described as an odds ratio, we think immediately of the disease odds ratio, which is the ratio of disease odds in the exposed group, a/b, divided by the disease odds in the unexposed group, c/d. Case-control studies cannot directly estimate the disease odds ratio, however, because we have not sampled everyone in column $Y = 0$ and therefore do not know b and d, only $b \times f$ and $d \times f$. What they can calculate is an exposure odds ratio: the odds of exposure in the cases, a/c, divided by the odds of exposure in the noncases, $b \times f/d \times f = b/d$. In this design, the exposure odds ratio, which we refer to as a case/noncase exposure odds ratio, is equal to the disease odds ratio:

$$\text{Case/noncase exposure OR} = \frac{a/c}{bf/df} = \frac{ad}{bc} \times \frac{f}{f} = \frac{ad}{bc} = \frac{a/b}{c/d} = \text{Disease OR}$$

We can repeat this procedure with the case-cohort design where we sample from the "Total" column in Table 1, again, with a sampling fraction f. We can calculate an exposure odds ratio but this time with a different denominator. We divide the odds of exposure in the cases (a/c) by the odds of exposure in the total population $([(c + d) \times f]/[(a+b) \times f])$. Note that it is not possible for the case/noncase exposure odds ratio from the previous paragraph, $\frac{a/c}{b/d}$, to be equal to this exposure odds ratio $\frac{a/c}{(a+b)/(c+d)}$, which we will call the case-cohort exposure odds ratio. Therefore, even though we have followed the same statistical procedure as in the previous paragraph, the epidemiologic reality of this estimate is different. The case-cohort exposure odds ratio is not equal to the disease odds ratio but the risk ratio:

$$\text{Case/cohort exposure OR} = \frac{a/c}{(a + b)f/(c + d)f} = \frac{a(c + d)}{c(a + b)} \times \frac{f}{f} = \frac{a(c + d)}{c(a + b)} = \frac{a/(a + b)}{c/(c + d)} = \text{RR}$$

Note that the exposure odds ratio from a case-cohort design is not an approximation of the risk ratio. It is, in fact, a mathematically equivalent way of expressing the risk ratio. Here we see the mismatch between statistical interpretation and epidemiologic reality. Although we have used only 2×2 tables to this point, we could also use logistic regression to analyze our case-cohort study. We are taught that the exponentiated coefficient from a logistic regression must be interpreted as an odds ratio. When data from a case-cohort design are analyzed with logistic regression, the exponentiated coefficient cannot be interpreted as an odds ratio but only as a risk ratio.

The same is true for designs that use sample person-time rather than participants. In these designs, the denominator of the exposure odds ratio is the odds of exposure across a sample f of all person-time: $PT_1 \times f/PT_0 \times f = PT_1/PT_0$. This case/person-time exposure odds ratio is equal to the incidence rate ratio:

$$\text{Case/person} - \text{time exposure OR} = \frac{a/c}{fPT_1/fPT_0} = \frac{aPT_0}{cPT_1} \times \frac{f}{f} = \frac{aPT_0}{cPT_1} = \frac{a/PT_1}{c/PT_0} = \text{IRR}$$

Again, the case/person-time exposure odds ratio is not equal to the disease odds ratio, and yet most studies with designs that estimate the incidence rate ratio report odds ratios (1). Either they believe these designs can be interpreted as disease odds ratios or are choosing to report exposure odds ratios. Furthermore, calling it an odds ratio not only defeats the purpose of using these designs but can be misleading to the reader who assumes, reasonably, that when the term odds ratio is used without qualifier, it refers to the disease odds ratio.

The same logic as for case-control studies with incidence-density sampling can be applied to open cohorts that either match on time or are conducted in populations where the prevalence of exposure is constant and that are therefore sampling person-time (2). Estimates from these designs can be interpreted only as incidence rate ratios and not odds ratios. This is important because, according to Knol et al. (1), these designs are the most common case-control study design and also appear to be the most often misinterpreted.

WHY IS THIS IMPORTANT?

If the estimate from case-cohort or incidence-density sampling designs is not equal to the disease odds ratio, why do studies employing these designs continue to refer to their estimates as odds ratios? An even simpler way of thinking about this is that a risk ratio or incidence rate ratio cannot be equal to a disease odds ratio (unless all are equal to 1). Therefore, it is possible for an estimate to have, at most, only one of these interpretations. Using the case-cohort design and referring to the estimate as an odds ratio is equivalent to using a model that estimates risk ratios (e.g., log-binomial regression) and calling the parameter an odds ratio.

The literature on case-control studies sometimes uses confusing language that can lead to some of these misconceptions. One textbook, referring to a case-cohort design, states that (8, p. 84) "the expected EOR [exposure odds ratio] from this case-control study would closely approximate the risk ratio from a corresponding follow-up study, even if the follow-up study was never done!" In fact, such a design does more than approximate the risk ratio, it is an estimator of the risk ratio. Another popular textbook states, "relative risks cannot be calculated directly from a case-control study," because case-control studies obtain only an "estimate of relative risks based on the odds ratios that are obtained in the case-control studies" (5, p. 208). Again, this is not correct. Relative risks can be directly calculated from case-cohort designs, and this does not rely on any special relationship between the risk ratio and the disease odds ratio. The risk ratio and the case-cohort exposure odds ratio are mathematically equivalent. Even statements such as "using a case-cohort design, one can estimate the risk ratio" (7, p. 124) are potentially ambiguous given that the reader might think an alternative parameter can also be presented.

We wish to point out that we are not advocating for the use of any of the terms we are using here except for teaching. These terms are simply to point out that there are 3 different types of exposure odds ratio that do not share the same properties, and it is therefore incorrect to assume that all exposure odds ratios can be interpreted as a disease odds ratio.

HAVE THINGS IMPROVED?

We performed a modified version of the review by Knol et al. (1), selecting the 2 most-cited case-control studies in the past 5 years from each of the following journals: *Lancet*, *New England Journal of Medicine*, *Journal of the American Medical Association*, *British Medical Journal*, *Annals of Internal Medicine*, *American Journal of Epidemiology*, *International Journal of Epidemiology*, *Epidemiology*, *European Journal of Epidemiology*, and the *Journal of Epidemiology and Community Health*. The search strategy and related code to run the search can be found in Web Appendix 1 (available at https://academic.oup.com/aje), as well as a table of the parameters reported.

Of the 20 studies we reviewed, 19 reported an odds ratio and 1 reported a hazard ratio (Web Table 1). The latter was a nested case-control study with a known sampling fraction, allowing the authors to analyze their case-control data as though it were a cohort. Furthermore, 13 studies used a design that realistically estimated the incidence rate ratio, and only 4 used designs that estimated odds ratios. In some studies, ambiguity in the description of control selection made it difficult to determine which parameter was being estimated.

Two studies (9, 10) explicitly mentioned that their sampling design allows them to interpret their estimates as incidence rate ratios yet reported odds ratios as their main parameters. For example, Friis et al. (10, p. 349) state, "With the nested case–control design and risk set sampling of control participants, the ORs provide unbiased estimates of the corresponding incidence rate ratios in the underlying source population, without distortion by competing risks." Despite this awareness, the authors present odds ratios as their main parameter. As we have argued, this reflects a long-standing misconception about case-control studies: Rather than providing authors an option of whether to report an odds ratio or incidence rate ratio (or risk ratio as the case may be), the study design and in particular the sampling strategy for the controls directly determines what parameter is being estimated.

CONCLUSION

Many epidemiologists before us have laid out proofs and explanations for why some case-control study designs can be interpreted as risk ratios and incidence rate ratios (1–3). What has been missing from the literature and textbooks is the clarification that these study designs not only can be interpreted as risk ratios or incidence rate ratios but must be interpreted as such. Lack of understanding of this point can be clearly seen in the literature where odds ratios are reported as the parameter of interest regardless of the design. Furthermore, a clear definition of the different types of exposure odds has been lacking. The term exposure odds ratio should not be used without being clear about who is in the denominator. Finally, it is important to know that in a case-control study, the sampling strategy determines which measure of association you are estimating and should be reported clearly.

ACKNOWLEDGMENTS

Author affiliation: Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands (Jeremy A. Labrecque, Myriam M. G. Hunink, M. Arfan Ikram, M. Kamran Ikram).

M.M.G.H. receives royalties from Cambridge University Press for a textbook on medical decision making, reimbursement of expenses from the European Society of Radiology (ESR) for work on the ESR guidelines for imaging referrals, reimbursement of expenses from the European Institute for Biomedical Imaging Research (EIBIR) for membership of the Scientific Advisory Board, and research funding from the American Diabetes Association, the Netherlands Organization for Health Research and Development, the German Innovation Fund, and the Society for Medical Decision Making. The other authors report no conflicts.

REFERENCES

1. Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol*. 2008;168(9):1073–1081.
2. Vandenbroucke JP, Pearce N. Case-control studies: basic concepts. *Int J Epidemiol*. 2012;41(5):1480–1489.
3. Miettinen OS. Estimability and estimation in case-referent studies. *Am J Epidemiol*. 1976;103(2):226–235.

4. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol*. 1982; 116(3):547–553.
5. Gordis L. *Epidemiology*. 4th ed. Philadelphia, PA: Elsevier/Saunders; 2009.
6. Kestenbaum B, Adeney KL, Weiss NS, et al. *Epidemiology and Biostatistics: An Introduction to Clinical Research*. Basel, Switzerland: Springer International; 2009.
7. Rothman KJ, Greenland S, Lash T. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Williams; 2008.
8. Kleinbaum DG, Sullivan KM, Barker ND. *A Pocket Guide to Epidemiology*. New York, NY: Springer-Verlag New York; 2007.
9. Wiese AD, Griffin MR, Schaffner W, et al. Opioid analgesic use and risk for invasive pneumococcal diseases: a nested case-control study. *Ann Intern Med*. 2018;168(6): 396–404.
10. Friis S, Riis AH, Erichsen R, et al. Low-dose aspirin or nonsteroidal anti-inflammatory drug use and colorectal cancer risk: a population-based, case-control study. *Ann Intern Med*. 2015;163(5):347–355.