The case-control study

It is not uncommon to develop technology without a full understanding of the underlying science. Cell phones are manufactured daily even though nobody knows whether they are made of elementary particles, of vibrating strings, or of something else. My computer is not floating in the air, and nobody knows if gravitons are part of the explanation. If something works then it works, and how someone got it to work is of lesser importance for the user.

Not so in science. When a research method is proposed—be it a new measurement or a new study design—an explicit rationale is expected, because no external standard can tell that a method is working—that is, helping to discover the secrets of Nature. All that can be done is to explain why some method serves the purpose of science and leave room for a healthy debate.[1,2] To estimate an effect, for example, we compare the frequency of the outcome between two values of the exposure variable because that comparison may tell us about the exposure effect. To remove confounding bias, we condition on a confounder because mathematical ideas tell us that conditioning will block a confounding path.

The case-control design is a time-honored exception. When that method showed up in science at the beginning of the twentieth century,[3] no one had a solid explanation of why we may learn about cause-and-effect from a case-control study, much less why the exposure odds ratio estimates the exposure effect on the outcome. On the contrary, common sense tells us that a causal connection may be discovered by comparing the distribution of the outcome variable across the values of the causal variable, not by comparing the distribution of the causal variable in two strata of the outcome: "cases" and "controls".

Historical explanations for the case-control design—some of which are still echoed in textbooks and courses—have been sketchy and weak. Take, for example, the following reasoning which was offered by the authors of one of the first case-control studies:

"We feel that any study of the habits of individuals with cancer is of little value without a similar study of individuals without cancer. To know that a large percentage of patients with cancer have certain habits is of little value for inference unless we know what percentage of the community at large has the same habit." (Quoted in an article by Paneth et al[3]).

In some minds the authors of this text "provided a rationale for the use of controls in words hard to improve upon".[3]

Hard to improve upon?

First, the paragraph contains a contradiction. If we also need to know the distribution of the same habit in "the community at large" (second sentence), why do we need "a similar study of individuals without cancer" (first sentence)? The community at large includes people with cancer too; it is not synonymous with "individuals without cancer". Second, the text does not explain why it is helpful to study the habits of individuals with cancer in the first place, rather than study the occurrence of cancer in individuals with different habits. Third, we don't need to know that the frequency of the exposure in cases is large; only that it differs from the frequency in controls. As usual, pioneering ideas courageously pave a road in the dark, but they are similar to a first draft of a manuscript—untidy and imprecise. Easy to improve upon.

Many years later, we find three attempts to explain the logic of the case-control design: The first offers no explanation at all; the second explains the design as a sample from a cohort; and the third is anchored in the theorems of a causal diagram. I will name them, respectively, the retrospective story, the cohort story, and the diagram story.

The retrospective story

The retrospective story tells us that there are two ways to estimate an effect: a prospective study, where we follow exposed and unexposed into their future, and a retrospective study where we select people who already have the disease (cases) and people who don't (controls) and look back at their past exposure. There is no explanation, however, of why "looking back" is a method to learn about a future effect.

Moreover, if "looking back" at past exposure is as valid as "looking forward" at future disease, why are there constraints on the measure of effect we may compute from a case-control study? Why should we compute the exposure odds ratio rather than the exposure probability ratio or the exposure probability difference?

The retrospective story is circular reasoning, telling us that the case-control design is valid because it is a valid design. No critical mind should have accepted it, but that's the prevailing "explanation" in many basic courses and introductory textbooks of epidemiology.

The cohort story

The following text is an attempt to explain the case-control study as a sample from a cohort.

"[The case-control study] employs an extra step of sampling according to the outcome of individuals in the population. This extra sampling step can make a case-control study much more efficient than a cohort study of the entire population…"[4, page 73]

The story, as quoted above, is not quite accurate, however. Indeed, a case-control study may be viewed as a sample from some cohort – real or theoretical – which is based on disease status. But that cohort doesn't have to include an "entire population" (whatever that foggy term means). The source cohort in which a case-control study is nested may include *any specified group of people*—say, pilots of Singapore Airlines combined with left-handed residents of Los Angeles who like cats.

Following this premise, simple math shows how the exposure odds ratio may estimate the rate ratio, the probability ratio, or the disease odds ratio—depending on the method by which controls are sampled.[5] For example, when controls are sampled from some cohort at baseline (the so-called case-cohort design), the exposure odds ratio estimates the probability ratio. When controls are sampled from members of some cohort who remained disease-free (the cumulative design), the exposure odds ratio estimates the disease odds ratio.

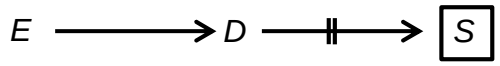
All questions about the validity of a particular case-control study turn into questions about the rules for sampling (selecting) cases and controls from the source cohort. Were they valid (unbiased)? And if not, can validity be restored at the analysis stage? For instance, matching controls to cases violates the requirement to randomly sample controls from the source cohort, but that violation may be rectified by stratification on the matched variables, which is equivalent to random sampling from each stratum of the source cohort.

The diagram story

The diagram story does not allude to any extra sampling step from a cohort study of some "entire population". The explanation is anchored in a causal diagram.

Figure 1 shows the basic causal structure of a case-control study: *E* is the exposure variable, *D* is the disease variable, and *S* is selection status. As usual, the effect of interest is *E*→*D*.

Figure 1. A basic causal diagram for a case-control study

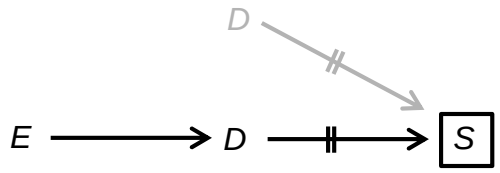


Disease status affects selection status because diseased people are over-sampled into a case-control study as compared with their disease-free counterparts. The box around *S* denotes conditioning, which here means restriction to one value of *S*. We always condition on selection status—not only in a case-control design[6]—because a study is obviously conducted only in the selected people. Following conditioning on *S*, the arrow *D*→*S* no longer contributes to any association (denoted by two lines over that arrow). In particular, *D* and *S* are dissociated.

Figure 1, however, does not tell the whole story. Several subtleties are revealed in the next series of diagrams.

Figure 2 shows that the arrow from *D* to *S* contributes to two causal paths: *E*→*D*→*S* and *D*→*S*. With this layout, it becomes clear that *E* and *D* collide at *S*, just like any two variables that share an effect. Since *E* collides through the same path as *D*, the structure is called uni-path colliding.[7] (Bi-path colliding describes the colliding of two causes through different paths.)

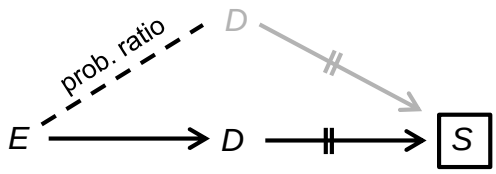
Figure 2. A different perspective of the causal diagram for a case-control study



In many circumstances, conditioning on a collider will result in colliding bias by adding a new, unwanted component to the association between the colliding variables.[7] That's true for uni-path colliding as well.

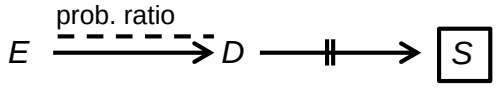
Specifically, if we try to estimate the effect *E*→*D* from a case-control study by most measures of association (e.g., the disease probability ratio) the estimator will contain colliding bias. The bias component is shown by a dashed line between *E* and *D* (Figure 3).

Figure 3. Colliding bias in a case-control study when estimating the disease probability ratio



Collapsing back to the layout of a single path, Figure 4 depicts the structure of uni-path colliding bias in a case-control study.

Figure 4. Colliding bias in a case-control study when estimating the disease probability ratio



Why does the bias arise?

Uni-path colliding bias arises because we alter the distribution of the effect variable, *S*, which would necessarily alter the distribution of its cause, *D*: Pr(*D*|*S*=1) ≠ Pr(*D*). Rare exceptions aside, if the probability distribution of *D* no longer arose from its causes alone, the conditional association between *D* and *E* does not reflect the effect *E*→*D*; bias must be present. Formally, the bias in the probability ratio may be written as follows:

Pr(*D*=1|*S*=1, *E*=1) / Pr(*D*=1|*S*=1, *E*=0) ≠
Pr(*D*=1|*E*=1) / Pr(*D*=1|*E*=0)

The proof of the inequality is given in Appendix A. The multiplier that restores equality is called "the bias factor".

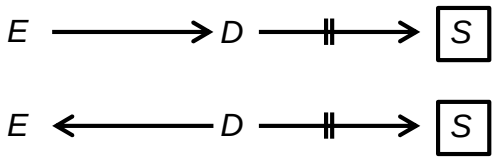
Pr(*D*=1|*S*=1, *E*=1) / Pr(*D*=1|*S*=1, *E*=0) =
Pr(*D*=1|*E*=1) / Pr(*D*=1|*E*=0) x **bias factor**

There is, however, at least one exception to the rule:

A measure of association between E and D that is indifferent to the direction of the arrow between the two variables will not be affected by conditioning on S. Specifically, uni-path colliding bias will not arise when the effect is estimated by an odds ratio.

A formal proof is given in Appendix B, but we may also get an intuitive explanation. Consider the following pair of diagrams (Figure 5), the top of which depicts a case-control study.

Figure 5. Two causal structures

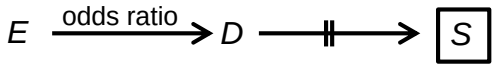


If the arrow between *E* and *D* is reversed (Figure 5, bottom diagram), conditioning on *S* will still alter the distribution of *D*, but no bias will be added when estimating the effect *E*←*D*. In causal inquiry the distribution of the cause of interest (here, *D*) may be altered with no penalty in the department of bias. For instance, exposed people are often over-sampled for an occupational cohort study.

If a measure of association is indifferent to the direction of the arrow between *E* and *D*, there is no difference between estimating the effect *E*→*D* (Figure 5, top diagram) and estimating the effect *E*←*D* (Figure 5, bottom diagram). The effect size is identical, which implies that the two diagrams are interchangeable. Since bias is absent from the bottom diagram, it must also be absent from the top diagram—for such a measure of association.

The odds ratio meets this condition. Whether *E*→*D* or *D*→*E*, the odds ratio is identical. Therefore, uni-path colliding bias is absent from a case-control study, so long as we estimate the effect by an odds ratio (Figure 6). Technically, that number may be called either the *exposure* odds ratio, or the *case-ness* odds ratio, or the *disease* odds ratio. Scientifically, only the latter term is meaningful. Moreover, the diagram story reveals that we may stop teaching students about the need to compute the exposure odds ratio from a case-control study. As far as uni-path colliding bias is concerned, the ratio of the odds of being a case in exposed to the odds of being a case in unexposed is an unbiased *disease* odds ratio.

Figure 6. No colliding bias in a case-control study when estimating the disease odds ratio



On the rare disease assumption

For reasons of efficiency and effort, a case-control study is preferred to a cohort study whenever the disease is rare. But for many years we have also heard that case-control studies *must* be confined to rare diseases. Is there any grain of truth in that claim?

The answer is neither simple nor short.

No matter how controls are sampled, we always compute an odds ratio. In the cumulative design, that odds ratio is the disease odds ratio, and neither the cohort story nor the diagram story calls for any "rare disease assumption". Other methods of control selection—case-cohort sampling or incidence density sampling—allow the odds ratio to estimate the probability ratio or the rate ratio, but no rare disease assumption is invoked, either.[5]

The assumption is needed only in the following situation: we use the classic, cumulative design *and claim to have estimated the probability ratio.* If the disease is rare in exposed and unexposed (say, frequency<0.15), the disease odds ratio and the disease probability ratio are approximately equal. (That approximation stems from similarity between the odds and the probability of a rare event—e.g., 0.1/0.9≈0.1).

But why might anyone want to claim that the disease odds ratio from the cumulative design estimate the disease probability ratio?

Well, most people hold the view that the probability ratio is the "correct" measure of effect, whereas the odds ratio is an aberrant measure that was forced upon us in the case-control design. (They also equate "risk" with "probability" and "relative risk" with "probability ratio".) Only a handful of authors, however, have written quasi-reasoning for that viewpoint, and the fundamental question of which measure of effect is preferred, if any, awaits a solid analysis. At any rate, all those who cannot tolerate the disease odds ratio—rightly or wrongly—must invoke the rare disease assumption for the classic, cumulative case-control design. Still, the assumption is never needed in case-cohort sampling or in incidence density sampling.

Acknowledgement: Doron Shahar - for comments on a draft manuscript and technical help.

References

1. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000:11;550-560
2. Shahar E, Shahar DJ. Marginal structural models: much ado about (almost) nothing. *Journal of Evaluation in Clinical Practice* 2013;19:214-22
3. Paneth N, Susser E, Susser M. Origins and early development of the case-control study: Part 2, The case-control study from Lane-Claypon to 1950. *Soz Praventivmed* 2002;47:359-65
4. Rothman KJ, Greenland S. *Modern Epidemiology* (second edition), Lippincott-Raven, 1998
5. Pearce N. What does the odds ratio estimate in a case-control study? *Int J Epidemiol* 1993:22;1189-92
6. Shahar E, Shahar DJ. Causal diagrams and the cross-sectional study. *Clinical Epidemiology* 2013;5:57–65
7. Shahar E, Shahar DJ. Causal diagrams and three pairs of biases. In: *Epidemiology – Current Perspectives on Research and Practice* (Lunet N, Editor). www.intechopen.com/books/epidemiology-current-perspectives-on-research-and-practice, 2012:pp. 31-62

Appendix A

The proof is based on the idea of conditional probability: $\Pr(A|B) = \Pr(A,B)/\Pr(B)$

$$\Pr(D=1|S=1,E=e) = \frac{\Pr(D=1,S=1,E=e)}{\Pr(S=1,E=e)}$$

$$= \frac{\Pr(S=1|D=1,E=e)\times\Pr(D=1,E=e)}{\Pr(S=1|E=e)\times\Pr(E=e)}$$

$$= \frac{\Pr(S=1|D=1,E=e)\times\Pr(D=1|\,E=e)\times\Pr(E=e)}{\Pr(S=1|E=e)\times\Pr(E=e)}$$

Given the structure $E\rightarrow D\rightarrow S$, the variables S and E are independent, conditional on D. Omitting E=e from Pr(S=1|D=1, E=e) and noting the cancellation of Pr(E=e), we get:

$$= \frac{\Pr(S=1|D=1)\times\Pr(D=1|\,E=e)}{\Pr(S=1|E=e)}$$

For E=1, we write:

$$\frac{\Pr(S=1|D=1)\times\Pr(D=1|\,E=1)}{\Pr(S=1|E=1)}$$

and for E=0, we write:

$$\frac{\Pr(S=1|D=1)\times\Pr(D=1|\,E=0)}{\Pr(S=1|E=0)}$$

Therefore, following conditioning on *S*, we estimate the effect *E*→*D* by the following probability ratio:

$$\frac{\Pr(S=1|D=1)\times\Pr(D=1|\,E=1)/\Pr(S=1|E=1)}{\Pr(S=1|D=1)\times\Pr(D=1|\,E=0)/\Pr(S=1|E=0)}$$

$$=\frac{\Pr(S=1|D=1)\times\Pr(D=1|E=1)\times\Pr(S=1|E=0)}{\Pr(S=1|D=1)\times\Pr(D=1|E=0)\times\Pr(S=1|E=1)}$$

$$=\frac{\Pr(D=1|E=1)}{\Pr(D=1|E=0)}\ \times\ \left[\frac{\Pr(S=1|E=1)}{\Pr(S=1|E=0)}\right]^{-1}$$

To summarize:

$$\frac{\Pr(D=1|S=1,E=1)}{\Pr(D=1|S=1,E=0)}$$

$$=\frac{\Pr(D=1|E=1)}{\Pr(D=1|E=0)}\ \times\ \left[\frac{\mathbf{Pr}(\boldsymbol{S=1|E=1})}{\mathbf{Pr}(\boldsymbol{S=1|E=0})}\right]^{-1}$$

The probability ratio Pr(*D*=1|*E*=1) / Pr(*D*=1|*E*=0)] quantifies the effect *E*→*D*, under the structure *E*→*D*→*S*. The expression in boldface may be called the bias factor due to conditioning on *S*. Notice that the bias factor quantifies the effect of *E* on *S*, and if *E* has a null effect on *S* (because *E*→*D* is null), conditioning on *S* does not add bias.

Appendix B

$$\mathrm{Odds}(D=1|S=1,E=e)=\frac{\Pr(D=1|S=1,E=e)}{\Pr(D=0|S=1,E=e)}$$

Based on Appendix A, we may substitute:

$$=\frac{\Pr(S=1|D=1)\ \times\ \Pr(D=1|E=e)/\Pr(S=1|E=e)}{\Pr(S=1|D=0)\ \times\ Pr(D=0|E=e)/\Pr(S=1|E=e)}$$

$$=\frac{\Pr(S=1|D=1)\ \times\ \Pr(D=1|E=e)}{\Pr(S=1|D=0)\ \times\ \Pr(D=0|E=e)}$$

For *E*=1, we write

$$\frac{\Pr(S=1|D=1)\ \times\ \Pr(D=1|E=1)}{\Pr(S=1|D=0)\ \times\ \Pr(D=0|E=1)}$$

and for *E*=0, we write:

$$\frac{\Pr(S=1|D=1)\ \times\ \Pr(D=1|E=0)}{\Pr(S=1|D=0)\ \times\ \Pr(D=0|E=0)}$$

The ratio of these two odds (the odds ratio) is reduced to the following:

$$\frac{\Pr(D=1|E=1)}{\Pr(D=0|E=1)}\Big/\frac{\Pr(D=1|E=0)}{\Pr(D=0|E=0)}$$

$$=\ \mathrm{Odds}(D=1|E=1)\ /\mathrm{Odds}(D=1|E=0)$$

which is the odds ratio for effect of *E* on *D*. Conditioning on *S* did not add bias.