Statistics and Probability Letters 125 (2017) 104–109

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

The directions of selection bias

Zhichao Jiang [a,*], Peng Ding [b]

[a] Department of Politics, Princeton University, NJ 08544, USA
[b] Department of Statistics, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 22 September 2016
Received in revised form 15 January 2017
Accepted 24 January 2017
Available online 13 February 2017

Keywords:
Monotonicity
Interaction
Odds ratio

ABSTRACT

We show that if the exposure and the outcome affect the selection indicator in the same direction and have non-positive interaction on the risk difference, risk ratio or odds ratio scale, the exposure-outcome odds ratio in the selected population is a lower bound for the true odds ratio.

1. Introduction

In epidemiology, observational studies are often used to investigate the relation between an exposure and a health outcome of interest. However, several potential biases might jeopardize our inference and conclusions (Greenland, 2005). Selection bias arises when the selected population is not representative of the target population of interest. As a consequence of selection bias, the association between exposure and outcome in the selected population differs from the association in the target population (Hernán et al., 2004).

In case-control studies, causal conclusions are more likely to be subject to selection bias than other epidemiologic studies (Geneletti et al., 2009). In a case-control study that recruits all (or most) of the diseased subjects and a small fraction of non-diseased subjects, the famous doctrine is that the selection of controls should not depend on their exposure status (Huang and Lee, 2015). Failing to satisfy this can lead to biased results. Previously, many researchers have discussed selection bias (e.g. Mezei and Kheifets, 2006). Some researchers derived the bias analytically (Nguyen et al., 2016), and some proposed methods to recover or adjust for selection bias (e.g. Bareinboim and Tian, 2014; Didelez et al., 2010; Yanagawa, 1984; Greenland, 2003; Valeri and Coull, 2016; Bareinboim and Pearl, 2012). We advance the literature by establishing qualitative relations between the exposure-outcome association in the selected population and that in the target population.

In this paper, we first consider the setting of the case-control studies with three variables (i.e., a binary exposure, a binary outcome and a binary indicator of selection), and then comment on the setting with covariates. Based on a decomposition of the odds ratio in the selected population, we show that if the exposure and the outcome affect the selection indicator in the same direction and have non-positive interaction on the risk ratio, odds ratio or risk difference scale, the odds ratio in the selected population is smaller than or equal to the true odds ratio in the target population. This relation can help us to draw qualitative conclusion about the true odds ratio. Compared with previous literature, we do not need prior quantitative knowledge of some unknown parameters, which are required in the sensitivity analysis and the adjustment methods. In contrast, we require some prior qualitative knowledge of the selection mechanism, and obtain the qualitative relation between the observed odds ratio and the true odds ratio.

* Corresponding author.
E-mail address: zhichaoj@princeton.edu (Z. Jiang).

http://dx.doi.org/10.1016/j.spl.2017.01.022
0167-7152/

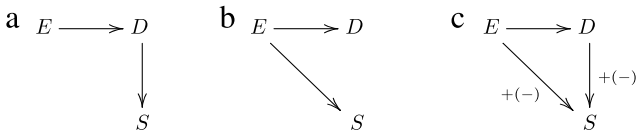

Fig. 1. Illustrative directed acyclic graphs.

2. Main results for the directions of selection bias for the odds ratio

We first introduce the notation. Let E be a binary exposure variable with $E = 1$ for treatment and $E = 0$ for control, and D be a binary outcome variable with $D = 1$ if disease is present and $D = 0$ otherwise. Let S be the binary indicator of selection with $S = 1$ if selected. For any binary variables A and B and a general variable C, we define

$$\begin{aligned}
\mathrm{OR}_{AB|C=c} &= \frac{P(A=1,B=1\mid C=c)P(A=0,B=0\mid C=c)}{P(A=1,B=0\mid C=c)P(A=0,B=1\mid C=c)},\\
\mathrm{RR}_{AB|C=c} &= \frac{P(B=1\mid A=1,C=c)}{P(B=1\mid A=0,C=c)},\\
\mathrm{RD}_{AB|C=c} &= P(B=1\mid A=1,C=c)-P(B=1\mid A=0,C=c),
\end{aligned}$$

as the odds ratio, risk ratio and risk difference of two random variables A and B conditional on $C = c$, respectively. For simplicity, we consider the setting without covariates and comment on the setting with covariates later. We are concerned about the true odds ratio, OR_{ED}, in the target population. However, from the selected population, we can estimate only the odds ratio conditional on $S = 1$, $\mathrm{OR}_{ED|S=1}$. In general, OR_{ED} and $\mathrm{OR}_{ED|S=1}$ are different, and they are related by an interaction measure between E and D on S. On the risk ratio scale, the multiplicative interaction of exposure and outcome on the selection indicator (VanderWeele, 2015) is defined as

$$\mathrm{Inter}_{\mathrm{RR}} = \frac{P(S=1\mid D=1,E=1)P(S=1\mid D=0,E=0)}{P(S=1\mid D=1,E=0)P(S=1\mid D=0,E=1)}.$$

The following result shows a well known relation between $\mathrm{OR}_{ED|S=1}$ and OR_{ED} (Kleinbaum et al., 1982; Greenland, 1996; Rothman et al., 2008; Greenland, 2009).

Result 0. We have

$$\mathrm{OR}_{ED|S=1} = \mathrm{OR}_{ED} \times \mathrm{Inter}_{\mathrm{RR}}. \tag{1}$$

Formula (1) states that the odds ratio in the selected population equals the true odds ratio multiplied by the interaction, on the risk ratio scale, of the exposure and outcome on the selection indicator.

Berkson (1946) gave numerical examples to show that the association between two diseases in the hospital population (selected population) is unrepresentative of that in the target population. In his examples, the two diseases are independent in the target population, but are positively associated in the selected population. With some abuse of notation, we let E and D indicate the occurrences of the two diseases respectively. Because E and D are independent in the target population, $\mathrm{OR}_{ED} = 1$, and thus according to (1), $\mathrm{OR}_{ED|S=1} = \mathrm{Inter}_{\mathrm{RR}}$, i.e., the odds ratio in the selected population equals the multiplicative interaction of E and D on selection. Berkson's choices of selection probabilities make $\mathrm{Inter}_{\mathrm{RR}} > 1$, which results in positive associations between E and D in the selected population. Note that the relation $\mathrm{OR}_{ED|S=1} = \mathrm{Inter}_{\mathrm{RR}}$ is also the fundamental identity in case-only designs for identifying gene–environment interactions (Piegorsch et al., 1994; Yang et al., 1999). For more discussion on collider bias, see Ding and Vanderweele (2016) and Ding et al. (2017).

If $P(S = 1 \mid D = d, E = e)$ is constant in d or e, then $\mathrm{Inter}_{\mathrm{RR}} = 1$ and hence $\mathrm{OR}_{ED|S=1} = \mathrm{OR}_{ED}$. This is related to the collapsibility conditions for the odds ratio (Didelez et al., 2010; Bareinboim and Pearl, 2012; Whittemore, 1978; Guo and Geng, 1995; Xie et al., 2008), i.e., if $D \perp\!\!\!\perp S \mid E$ or $E \perp\!\!\!\perp S \mid D$, then $\mathrm{OR}_{ED|S=s} = \mathrm{OR}_{ED}$ for $s = 0, 1$.

Therefore, the odds ratio in the selected population will be equal to the odds ratio in the target population under either of the following two scenarios: (a) the probability of being selected is dependent only on the subjects' outcome status, but the exposure does not directly affect the subjects' selection or inclusion probabilities (Fig. 1(a)); (b) the probability of being selected is dependent only on the subjects' exposure status, but the outcome does not directly affect the subjects' selection or inclusion probabilities (Fig. 1(b)). If the study recruits all of the diseased subjects as cases, and the selection of non-diseased subjects is independent of their exposure status, then condition (a) holds because $P(S = 1 \mid D = 1, E = e) = 1$ and $S \perp\!\!\!\perp E \mid D = 0$. Thus, the odds ratio in the selected population equals the odds ratio in the target population, which justifies the doctrine mentioned in Section 1.

If the collapsibility conditions, $D \perp\!\!\!\perp S \mid E$ and $E \perp\!\!\!\perp S \mid D$, do not hold but there is no interaction of E and D on S on the risk ratio scale, we still have $\mathrm{Inter}_{\mathrm{RR}} = 1$, which immediately gives the following result.

Result 1. If there is no interaction of E and D on S on the risk ratio scale, i.e., $\mathrm{Inter}_{\mathrm{RR}} = 1$, then $\mathrm{OR}_{ED|S=1} = \mathrm{OR}_{ED}$.