Non-differential Outcome Misclassification

Tetsuro Oda

History

Created - 4, Jan, 2023.

What does this note answer for you?

This note gives a bit more detail for Chubak et al. (2012) and Newcomer et al. (2019a) with regard to **non-differential outcome misclassification in RR**, by doing some easy maths.

Specifically, this note would answer the following questions (hopefully without errors) in a situation where you want to estimate RR for a binary exposure under non-differential outcome misclassification:

- Why is it true that imperfect sensitivity with perfect specificity does not introduce bias?
- Why is it true that perfect sensitivity with imperfect specificity DOES introduce bias?
- Why is the bias towards the null?
- Why is the bias larger when the true RR is larger?
- Why is the bias larger when the outcome is rarer?
- Why is the bias larger when the exposure is more common?
- How would you derive the correction formula for RR using only PPVs stratified by exposure?
- Why is the perfect specificity insufficient to guarantee zero bias for RD, OR, and HR in a similar setting?
- How would you do quantitative bias analysis? What is it like?

Contents

Understand non-differential outcome misclassification in RR

Introducing some notations

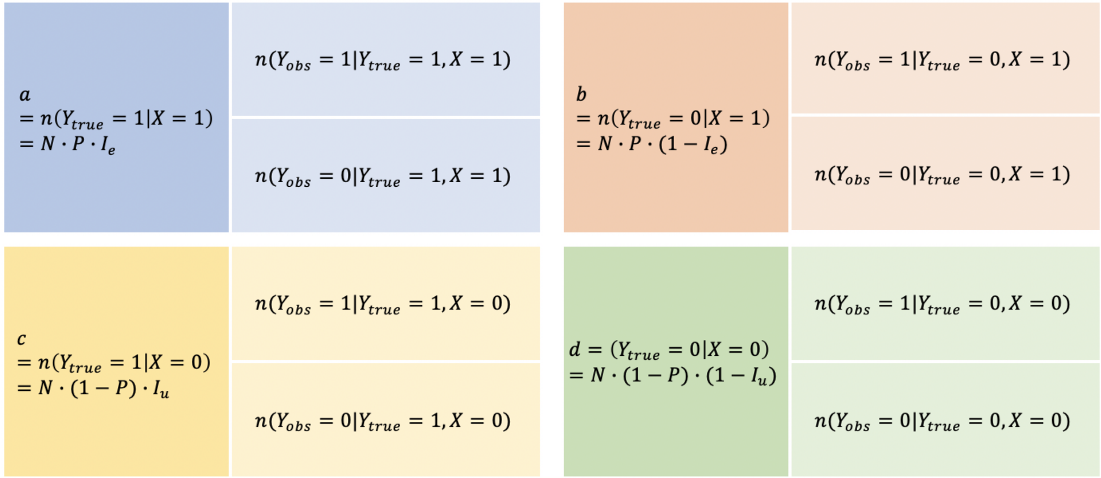
Let me first organise some notations. The following table represents a true 2 by 2 table without outcome misclassification. See the left side of Table 1. Now, can we denote a 2 by 2 table with outcome misclassification on the right, using the notation in the true 2 by 2 table? The answer is yes, and the method is clearly stated in Table 4 of Appendix in Chubak et al. (2012), but I will elaborate it a bit more in detail.

Table 1: True (left) and observed (right) 2 by 2 table

	Y_{true}			Y_{obs}		
	$Y_{true} = 1$	$Y_{true} = 0$	Total	$Y_{obs} = 1$	$Y_{obs} = 0$	Total
$X = 1$	a	b	$a + b$	a'	b'	$a' + b'$
$X = 0$	c	d	$c + d$	c'	d'	$c' + d'$

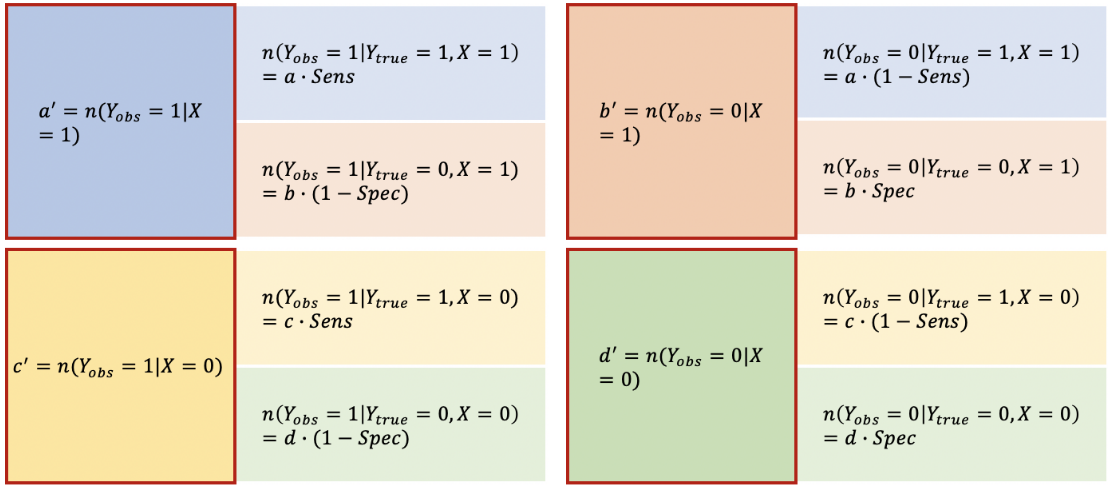
To do so, let's split each of the four cells in the true 2 by2 table into two quantities by an outcome definition as shown in Table 2, where $Y_{true} = 1$ means true cases, $Y_{obs} = 1$ means observed cases that meet an outcome definition, $X = 1$ means exposed. $n()$ is a symbol expressing the number of units that meet the condition. Also, N, P, I_e, I_u denote the number of all units, the proportion of exposeds, the incidence in the exposeds, the incidence in the unexposeds, respectively.

Table 2: Different representation of Table 1. Four cells stratified by Y_{true} to eight cell counts



a $= n(Y_{true} = 1 \vert X = 1)$ $= N \cdot P \cdot I_e$	$n(Y_{obs} = 1 \vert Y_{true} = 1, X = 1)$	b $= n(Y_{true} = 0 \vert X = 1)$ $= N \cdot P \cdot (1 - I_e)$	$n(Y_{obs} = 1 \vert Y_{true} = 0, X = 1)$
	$n(Y_{obs} = 0 \vert Y_{true} = 1, X = 1)$		$n(Y_{obs} = 0 \vert Y_{true} = 0, X = 1)$
c $= n(Y_{true} = 1 \vert X = 0)$ $= N \cdot (1 - P) \cdot I_u$	$n(Y_{obs} = 1 \vert Y_{true} = 1, X = 0)$	$d = (Y_{true} = 0 \vert X = 0)$ $= N \cdot (1 - P) \cdot (1 - I_u)$	$n(Y_{obs} = 1 \vert Y_{true} = 0, X = 0)$
	$n(Y_{obs} = 0 \vert Y_{true} = 1, X = 0)$		$n(Y_{obs} = 0 \vert Y_{true} = 0, X = 0)$

Table 3: Different representation of Table 3. The colours match up between Table 3 and 4 for ease of understanding.



$a' = n(Y_{obs} = 1 \vert X$ $= 1)$	$n(Y_{obs} = 1 \vert Y_{true} = 1, X = 1)$ $= a \cdot Sens$	$b' = n(Y_{obs} = 0 \vert X$ $= 1)$	$n(Y_{obs} = 0 \vert Y_{true} = 1, X = 1)$ $= a \cdot (1 - Sens)$
	$n(Y_{obs} = 1 \vert Y_{true} = 0, X = 1)$ $= b \cdot (1 - Spec)$		$n(Y_{obs} = 0 \vert Y_{true} = 0, X = 1)$ $= b \cdot Spec$
$c' = n(Y_{obs} = 1 \vert X = 0)$	$n(Y_{obs} = 1 \vert Y_{true} = 1, X = 0)$ $= c \cdot Sens$	$d' = n(Y_{obs} = 0 \vert X$ $= 0)$	$n(Y_{obs} = 0 \vert Y_{true} = 1, X = 0)$ $= c \cdot (1 - Sens)$
	$n(Y_{obs} = 1 \vert Y_{true} = 0, X = 0)$ $= d \cdot (1 - Spec)$		$n(Y_{obs} = 0 \vert Y_{true} = 0, X = 0)$ $= d \cdot Spec$

Using these notations in Table 2 as well as sensitivity and specificity of outcome definition, we can connect true cell counts with observed cell counts, as shown in Table 3. Note that the colours between Table 2 and 3 match up so that you can see the relationship easily. In this representation, **outcome misclassification is considered to be non-differential by exposure, hence there is only one fixed symbol for sensitivity and specificity, respectively, irrespective of the exposure status**.

Using the notations in Table 3, the observed risk ratio is:

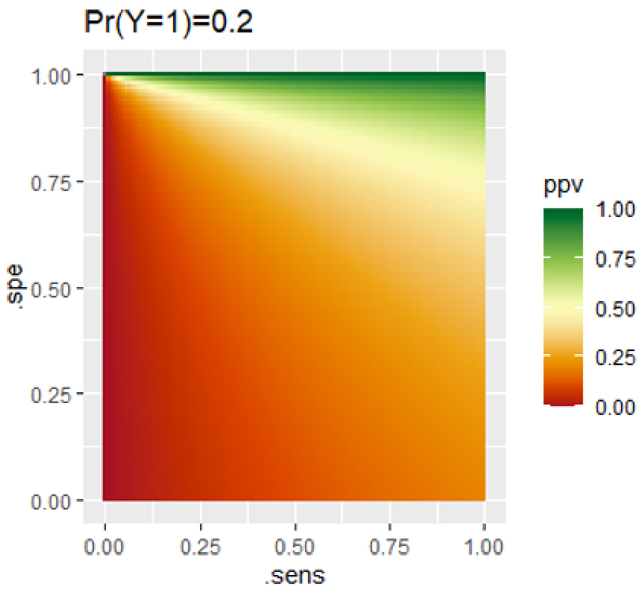
$$RR_{obs} = \frac{a'/(a'+b')}{c'/(c'+d')} = \frac{(a \cdot Sens + b \cdot (1 - Spec))/(a+b)}{(c \cdot Sens + d \cdot (1 - Spec))/(c+d)}.$$

As you can see, the denominators of both risks do not change, therefore, we only need to consider the numerators in thinking of bias introduced by outcome misclassification. Further, we can get a relationship between $RR_{true} = \frac{a/(a+b)}{c/(c+d)}$ and RR_{obs} as follows:

$$RR_{obs} = RR_{true} \cdot \frac{1 + b \cdot (1 - Spec)/(a \cdot Sens)}{1 + d \cdot (1 - Spec)/(c \cdot Sens)}$$

Note that the formula above is useful for the subsequent discussion to better understand how outcome misclassification occurs for RR but is NOT useful to back-calculate the true RR because specificity typically is not available in a validation study. Yet, we can visualise the relationship among sensitivity, specificity, and PPV given a prevalence of outcome, therefore, we could get a possible range of specificity without a validated specificity value, as shown in the following figure.

Figure1: PPV Heatmap with varying sensitivity and specificity.



How sensitivity and specificity of outcome definition bias RR

The biggest surprise to me when I first read Chubak et al. (2012) is that **as long as misclassification is non-differential between the exposure status, imperfect sensitivity does NOT bias the result when specificity is perfect**. Why is that? It is because when specificity is forced to be 1 in the equation above, the terms multiplied by $(1 - Spec)$ becomes zero and the sensitivity terms cancel out. In contrast, **even when sensitivity is perfect, the bias multiplier does not go to 1 with imperfect specificity** because the b/a and d/c terms.

What about the direction of bias? It seems to me that the bias is bidirectional, but these papers state that **bias in this setting is typically towards the null**. Let's have a closer look at the formula above, The bias direction should be determined by the ratio of the second terms in both numerator and denominator, so pick up and rearrange it:

$$\frac{b \cdot (1 - Spec)/(a \cdot Sens)}{d \cdot (1 - Spec)/(c \cdot Sens)} = \frac{b/a}{d/c} = \frac{1}{OR}.$$

Wow, we got the reciprocal of odds ratio for the outcome between exposed and unexposed. This means that if the true odds ratio is greater than 1, the bias will work in a way that decreases our estimate while our estimate will be biassed upward if the true OR is smaller than 1. In short, this result coincides with what the papers say; bias will be working always towards the null in this setting. **What setting? Non differential outcome misclassification and binary exposure**. **Conversely, there are cases where bias away from the null can occur in different settings, e.g., more than two or continuous exposure, differential misclassification, and etc.** See the two papers for detail.

Chubak et al. add some more insights. First, **they say that the more away from the null the true RR is, the more difference the observed Risk gets**. This is easily understood because the sensitivity and specificity of the outcome definition influences the true RR in a multiplicative manner. Another thing the paper notes is that **the more common the outcome is the less bias the observed RR gets**. Think about a super prevalent outcome with a very high I_e and I_u. Then b/a and d/c terms become closer to 0, meaning that the whole multiplier besides the true RR gets closer to 1. This also means that specificity itself is not influential as much as it is with a less prevalent outcome. Note that when an outcome is super prevalent, PPV is likely to be very high irrespective of specificity because of a very small $(b+d)$. Finally, though it does not seem to be stated in Chubak et al, **the bias seems to become larger when the proportion of exposure is larger** because the size of the exposed group only changes the numerator.

As a side note, which do you think is the most problematic miscount, false positive (FP) or false negative (FN)? The answer is FP because FP is determined by $(1 - Spec)$. The more FP we get with a common specificity between exposure status, the risks in both groups are diluted, therefore, the RR gets diluted as well.

Additional comments

Even when differential misclassification occurs, I was thinking that an RR would be unbiased if covariates associated with sensitivity and specificity are adjusted for, but it might be better not to rely on such an assumption and to do bias analysis.

Chance alone might cause differential misclassification in our specific sample, so do not forget about it.

Table 5: 2 by 2 table classifying true Y against observed Y.

	$Y_{true} = 1$	$Y_{true} = 0$
$Y_{obs} = 1$	$TP = (a+c) \cdot Sens$	$FP = (b+d) \cdot (1 - Spec)$
$Y_{obs} = 0$	$TN = (a+c) \cdot (1 - Sens)$	$TN = (b+d) \cdot Spec$
Total	$a + c$	$b + d$

Correction for misclassified RR using PPVs

Up until now, we have looked into how non-differential outcome misclassification changes RR estimates. In this section, I would like to derive Formula 3 in Newcomer et al. (2019), which is with a slightly different notation,

$$RR_{true} = RR_{observed} \cdot \frac{PPV_1}{PPV_0}$$

where PPV_1 is a positive predictive value in an exposed group and PPV_0 is that in an unexposed group, under a non-differential sensitivity assumption. Note that this formula is first given by Brenner & Gefeller (1993), and Newcomer et al. (2019) says this formula is a special case of Lash et al. (2011).

Let's derive this formula. First, express the true RR by the observed RR:

$$RR_{true} = \frac{Pr(Y_{true}=1|X=1)}{Pr(Y_{true}=1|X=0)} = \frac{Pr(Y_{obs}=1|X=1)}{Pr(Y_{obs}=1|X=0)} \cdot \frac{Pr(Y_{true}=1|X=1)/Pr(Y_{obs}=1|X=1)}{Pr(Y_{true}=1|X=0)/Pr(Y_{obs}=1|X=0)}$$
$$= RR_{obs} \cdot \frac{Pr(Y_{true}=1|X=1)Pr(Y_{obs}=1|X=0)}{Pr(Y_{true}=1|X=0)Pr(Y_{obs}=1|X=1)}.$$

Using the following relationship for the exposed group:

$$Pr(Y_{obs}=1|X=1)$$
$$= Pr(Y_{true}=1, Y_{obs}=1|X=1)/Pr(Y_{true}=1|Y_{obs}=1, X=1)$$
$$= Pr(Y_{true}=1, Y_{obs}=1|X=1)/PPV_1$$

and that for the unexposed group: $Pr(Y_{obs}=1|X=0) = Pr(Y_{true}=1, Y_{obs}=1|X=0)/PPV_0$, we get:

$$RR_{true} = RR_{obs} \cdot \frac{Pr(Y_{true}=1|X=1)(Pr(Y_{true}=1, Y_{obs}=1|X=0)/PPV_0)}{Pr(Y_{true}=1|X=0)(Pr(Y_{true}=1, Y_{obs}=1|X=1)/PPV_1)}$$
$$= RR_{obs} \cdot \frac{Pr(Y_{true}=1, Y_{obs}=1|X=0)/Pr(Y_{true}=1|X=0)}{Pr(Y_{true}=1, Y_{obs}=1|X=1)/Pr(Y_{true}=1|X=1)} \cdot \frac{PPV_1}{PPV_0}$$
$$= RR_{obs} \cdot \frac{Pr(Y_{obs}=1|Y_{true}=1, X=0)}{Pr(Y_{obs}=1|Y_{true}=1, X=1)} \cdot \frac{PPV_1}{PPV_0}$$
$$= RR_{obs} \cdot \frac{Sens_0}{Sens_1} \cdot \frac{PPV_1}{PPV_0}$$
$$= RR_{obs} \cdot \frac{PPV_1}{PPV_0}$$

The last equation holds when the non-differential sensitivity holds. This formula has a weaker assumption than the aforementioned formula, and we only need to know PPVs. Yet, we need to have that in the exposed and unexposed group separately. Considering a typical validation study, information on PPV_1 and PPV_0 is likely to be unavailable, therefore, it might still have a limited usefulness for a correction purpose.

Consideration for RD, OR

According to the two papers, other estimators like risk difference(RD), odds ratio (OR) are biassed even when specificity is perfect. Let's confirm that.

The difference between the observed and true risk is:

$$R_{obs} - R_{true} = \frac{a \cdot (Sens - 1) - b \cdot (1 - Spec)}{a + b} - \frac{c \cdot (Sens - 1) - d \cdot (1 - Spec)}{c + d}.$$

When specificity is 1, it reduces to the following and is still non-zero:

$$R_{obs} - R_{true} = \frac{a \cdot (Sens - 1)}{a + b} - \frac{c \cdot (Sens - 1)}{c + d},$$

which confirms that **RD is biassed even when specificity is perfect**.

What about OR? The observed OR is expressed as:

$$OR_{obs} = \frac{(a \cdot Sens + b \cdot (1 - Spec))(c \cdot (1 - Sens) + d \cdot Spec)}{(a \cdot (1 - Sens) + b \cdot Spec)(c \cdot Sens + d \cdot (1 - Spec))}.$$

When specificity is 1, the formula reduces to:

$$OR_{obs|Spec=1} = \frac{a \cdot Sens \cdot (c \cdot (1 - Sens) + d)}{(a \cdot (1 - Sens) + b) \cdot c \cdot Sens},$$

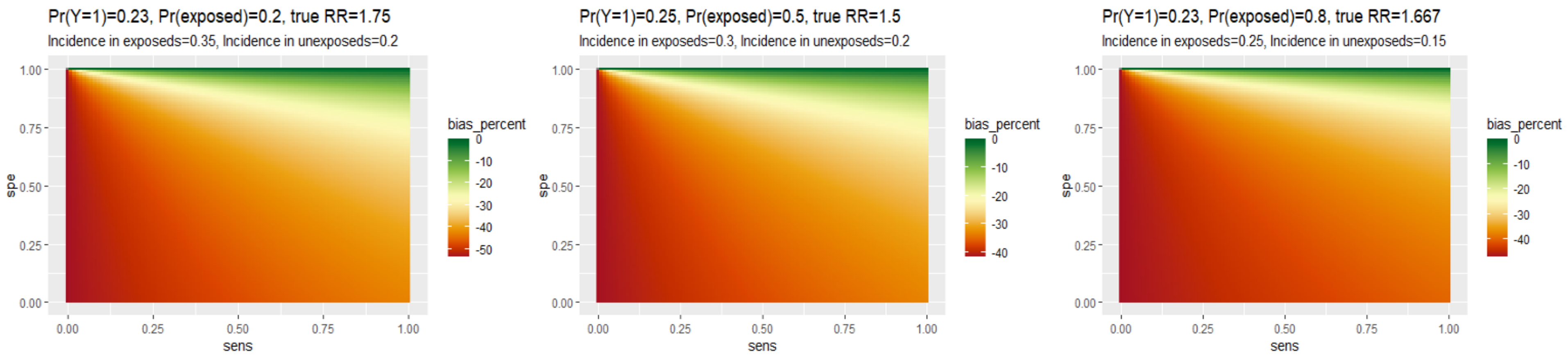
which is not equal to $OR_{true} = \frac{ad}{bc}$.

Quantitative Bias Analysis

Quantitative bias analysis is an analysis quantifying bias using simulation. It seems to me that it is always good to do QBA for study planning. By doing QBA, visualisation of how much bias we would get would be helpful even with a non-differential misclassification assumption because the two main formulas we have seen do not tell us how much bias we would get in a straightforward manner.

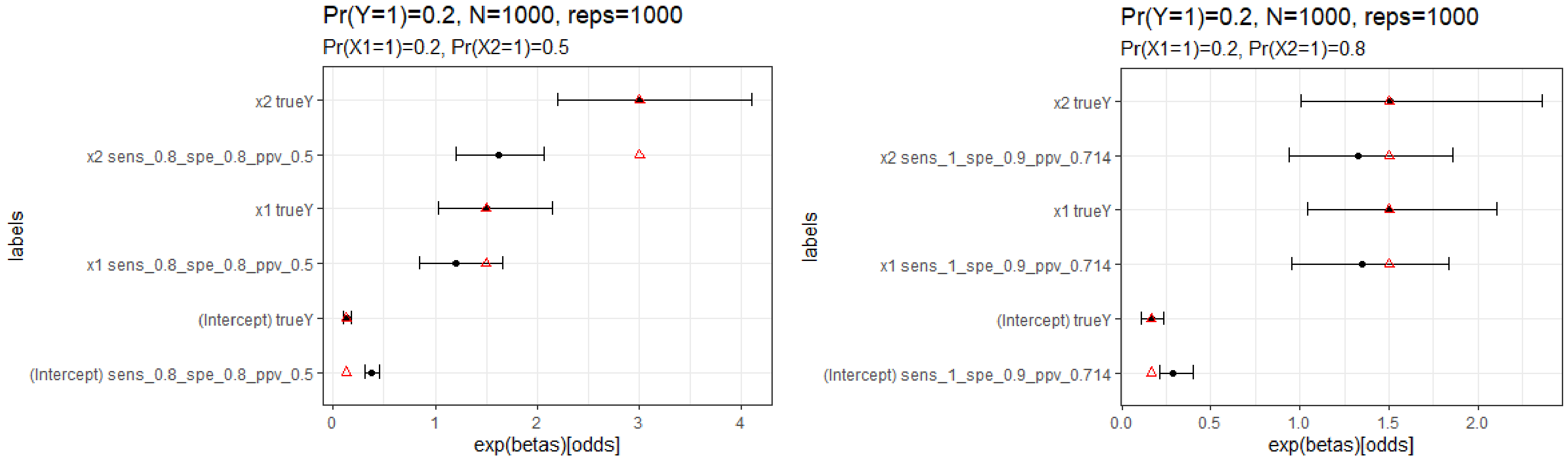
The three figures are some examples of visualisation of Chubak and colleagues' work for bias percents in non-differential outcome misclassification (Appendix Table 4). This way, we see that without very high specificity, we would get a strong bias towards the null. My code for this simulation is found here.

Figure 2: Heatmap of bias percent, expanding Appendix Table 4 in Chubak et al (2012).



Another simulation I did is QBA for non-differential outcome misclassification in multiple logistic regression. My code is found here. The example figures (Figure 3) all show that ORs are biassed towards the null.

Figure3: Estimated true and non-differentially misclassified ORs and 95% bootstrap confidence intervals from multiple logistic regression.



Reading list for myself

- Newcomer et al. (2019b): seems to be a good resource to do QBA for differential misclassification.
- Disease misclassification in electronic healthcare database studies: Deriving validity indices—A contribution from the ADVANCE project
- Outcome misclassification: Impact, usual practice in pharmacoepidemiology database studies and an online aid to correct biased estimates of risk ratio or cumulative incidence
- Mitigation of biases in estimating hazard ratios under non-sensitive and non-specific observation of outcomes–applications to influenza vaccine effectiveness
- Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: a systematic review