frontiers

Selection and Misclassification Biases in Longitudinal Studies

Denis Haine [1,3,*], **Ian Dohoo** [2,3] **and Simon Dufour** [1,3]

[1] *Faculté de médecine vétérinaire, Université de Montréal, St-Hyacinthe, QC, Canada*
[2] *Centre for Veterinary Epidemiological Research, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, PE, Canada*
[3] *Canadian Bovine Mastitis and Milk Quality Research Network, St-Hyacinthe, QC, Canada*

Correspondence*:
Simon Dufour
simon.dufour@umontreal.ca

ABSTRACT

Using imperfect tests may lead to biased estimates of disease frequency and measures of association. Many studies have looked into the effect of misclassification on statistical inferences. These evaluations were either within a cross-sectional study framework, assessing biased prevalence, or for cohort study designs, evaluating biased incidence rate or risk ratio estimates based on misclassification at one of the two time-points (initial assessment or follow-up). However, both observations at risk and incident cases can be wrongly identified in longitudinal studies, leading to selection and misclassification biases, respectively. The objective of this paper was to evaluate the relative impact of selection and misclassification biases resulting from misclassification, together, on measures of incidence and risk ratio.

To investigate impact on measure of disease frequency, data sets from a hypothetical cohort study with two samples collected one month apart were simulated and analyzed based on specific test and disease characteristics, with no elimination of disease during the sampling interval or clustering of observations. Direction and magnitude of bias due to selection, misclassification, and total bias was assessed for diagnostic test sensitivity and specificity ranging from 0.7 to 1.0 and 0.8 to 1.0, respectively, and for specific disease contexts, i.e. disease prevalences of 5 and 20%, and disease incidences of 0.01, 0.05, and 0.1 cases/animal-month. A hypothetical exposure with known strength of association was also generated. A total of 1,000 cohort studies of 1,000 observations each were simulated for these six disease contexts where the same diagnostic test was used to identify observations at risk at beginning of the cohort and incident cases at its end.

Our results indicated that the departure of the estimates of disease incidence and risk ratio from their true value were mainly a function of test specificity, and disease prevalence and incidence. The combination of the two biases, at baseline and follow-up, revealed the importance of a good to excellent specificity relative to sensitivity for the diagnostic test. Small divergence from perfect specificity extended quickly to disease incidence over-estimation as true prevalence increased and true incidence decreased. A highly sensitive test to exclude diseased subjects at baseline was of less importance to minimize bias than using a highly specific one at baseline.

Near perfect diagnostic test attributes were even more important to obtain a measure of association close to the true risk ratio, according to specific disease characteristics, especially its prevalence. Low prevalent and high incident disease lead to minimal bias if disease is diagnosed with high sensitivity and close to perfect specificity at baseline and follow-up. For more prevalent diseases we observed large risk ratio biases towards the null value, even with near perfect diagnosis.

Keywords: bias (epidemiology), longitudinal study, selection bias, misclassification, epidemiologic methods

1 INTRODUCTION

A cohort study is a longitudinal observational study in which a study population (i.e. a cohort) is selected and followed up in time (Dos Santos Silva, 1999; Rothman et al., 2012). Members of the cohort share a common experience (e.g. Kennel Club registered Labrador Retrievers born after January 1st, 2010; Clements et al., 2013) or condition (e.g. litters from *A. pleuropneumoniae* infected sows; Tobias et al., 2014). Two cohorts are often included in these longitudinal studies, one experiencing a putative causal event or condition (exposed cohort), and the other being an unexposed (reference) cohort. Cohort study is the standard study design to estimate the incidence of diseases and identify their natural history, by analyzing the association between a baseline exposure and risk of disease over the follow-up period. This type of study is characterized by the identification of a disease-free population (i.e. subjects with the outcome at baseline are excluded from the follow-up), and their exposure to a risk factor is assessed. The frequency of the outcome (generally the incidence of a disease or death) is measured and related to exposure status, expressed as a risk ratio (RR). Therefore it is assumed that prevalent and non-prevalent cases can be differentiated with no error so that only susceptible individuals are included in the cohort. Incident cases are likewise supposed to be correctly identified.

However, any measurement is prone to potential errors, as a result of subjective evaluations, imperfect diagnostic tests, reporting errors (deliberate or not), recall deficiencies, or clerical errors. Obtaining "error-free" measurements is a desirable objective but it is usually much more expensive to use "gold-standard" measurements, or they are simply not available, leaving the researcher with "less-than-ideal" measurement tools. Wrong classification at baseline and at follow-up are both misclassification biases, in the former the bias resulting from misclassification could be considered a selection bias, as the wrong (diseased) subjects are included in the cohort (Rothman et al., 2012) while in the latter, it would be commonly defined as misclassification bias (Delgado-Rodriguez and Llorca, 2004). Such errors of measurement or misclassification in exposure variables, outcomes or confounders can bias inferences drawn from the data collected, often substantially (Quade et al., 1980), or decrease the power of the study (Bross, 1954; White, 1986). Many studies have looked into the effect of misclassification on statistical inferences, including biased prevalence and incidence rate estimates (Rogan and Gladen, 1978; Quade et al., 1980) and biased relative risk estimates (Barron, 1977; Greenland, 1980). Nondifferential misclassification of disease leads in general to bias towards null in the estimated associations as well as reduced statistical efficiency (Bross, 1954; Barron, 1977; Copeland et al., 1977). This bias depends mainly on the specificity (Sp) of the test used (Copeland et al., 1977). If Sp of the test is perfect, then bias is absent (Poole, 1985). These evaluations were, however, either within a cross-sectional study framework, assessing biased prevalence, or for cohort study designs evaluating biased incidence rate or RR estimates but based on misclassification at only one of the two time-points (initial assessment or follow-up). However, both observations at risk and incident cases can be wrongly identified in longitudinal studies, leading to selection and misclassification biases, respectively.

The objective of this paper was to evaluate the relative impact of selection and misclassification biases resulting from misclassification, together, on measures of incidence and RR.

2 MATERIAL AND METHODS

To investigate the impact of concomitant selection and misclassification biases on measure of disease frequency, data sets from a hypothetical cohort study with two samples collected one time unit apart were simulated and analyzed based on specific test and disease characteristics, for a stable population over the follow-up time, and with no elimination of disease or clustering of observations. Direction and magnitude of bias due to selection, misclassification, and total bias was assessed for diagnostic test sensitivity (Se) and Sp ranging from 0.7 to 1.0 (0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.98, 0.99, 1) and 0.8 to 1.0 (0.8, 0.85, 0.9, 0.95, 0.98, 0.99, 1), respectively, and for specific disease contexts, i.e. disease prevalences of 5 and 20%, and disease incidences of 0.01, 0.05, and 0.1 cases/animal-time unit. The true case status (S_1) on first sample collection was used to identify observations at risk at the beginning of the cohort, while the second (S_2) was used to identify the true outcome. A hypothetical exposure with known strength of association (RR $\sim$3.0) was also generated. For demonstration purpose, simulations were also ran with a weaker RR of $\sim$1.5 (Supplementary Material). A total of 1,000 cohort studies of 1,000 observations each were simulated for these six disease contexts where the same diagnostic test was used to identify observations at risk at beginning of the cohort and incident cases at its end. On each datasets new S_1' and S_2' variables were generated by applying the scenario misclassification parameters to the S_1 and S_2 samples. Incidence and measures of association with the hypothetical exposure were then computed using first the S_1' and S_2' variables (total bias), then S_1' and S_2 (selection bias only), and finally the S_1 and S_2' variables (misclassification bias only).

Disease incidence was computed as the number of new cases at the end of the cohort divided by the number at risk at its beginning. Risk ratio was computed as the ratio of the risk of disease among observations who were exposed to the risk factor, to the risk among observations who were unexposed (Rothman et al., 2012). Data sets generation and estimation procedures were realized in R (R Core Team, 2017), and simulation code is available at `https://github.com/dhaine/cohortBias`.

3 RESULTS

Total biases resulting from selection and misclassification errors and according to given disease prevalence, Se, and Sp are illustrated for disease incidence and RR in Figures 1 and 2, respectively. These figures are contour plots where the lines are curves in the x, y-plane along which the function of the two variables on the vertical and horizontal axes (i.e. Se and Sp) has a constant value, i.e. a curve joins points of equal value (Courant et al., 1996). The true incidence rate (or RR) is therefore to be found at the upper right corner of the plot. For example, in the bottom left panel of Figure 1 the second line from the bottom is labelled 0.22. This line shows that, for a 5% disease prevalence and a true incidence rate of 0.1 case/animal-time unit, an apparent incidence estimate of 0.22 will be achieved by any combination of Sp and Se on this line (e.g. Sp = 0.845, Se = 0.7 or Sp = 0.87 and Se = 1.00). As an other example, in the upper right panel of this same figure, the first line at the top is labelled 0.02. It shows that, for a 5% disease prevalence and a true incidence rate of 0.01 case/animal-time unit, an apparent incidence estimate of 0.02 is achieved along this line by any combination of Se and Sp like, for example, a Sp of 1.00 and a Se of 0.955. The true incidence rate is given at the upper right corner, where Se and Sp are both 100%. Imperfect Se to identify individuals at risk at baseline and imperfect Sp to identify incident cases led to a mild under-estimation of the observed

disease incidence (Figures S1 and S2 in Supplementary Material). From these graphs we could also note that Sp has little effect on selection bias while Se has little effect on misclassification bias. Of the two, misclassification bias had a much bigger effect than selection bias. But overall, the combination of the two biases, at baseline and follow-up, revealed the importance of a good to excellent Sp relative to Se for the diagnostic test. Small divergence from perfect Sp extended quickly to disease incidence over-estimation as true prevalence increased and true incidence decreased (Figures 3 to 5). Selection and misclassification biases of a low prevalent and incident disease, diagnosed with close to perfect Sp, were minimal, reflecting the importance of choosing a highly specific test to improve identification of animal (or individual) unit at risk and incident case identification. The same effect was also observed with RR estimations (Figures S3 and S4). Similar results were found with a weaker exposure, RR of 1.5 (Figures S5 to S8).

4 DISCUSSION

Our results indicated that the departure of the estimates of disease incidence and risk ratio from their true value were mainly a function of test Sp, and disease prevalence and incidence. Imperfect Se to identify individuals at risk and imperfect Sp to identify incident cases led to a mild under-estimation of the observed disease incidence. The combination of the two biases, at baseline and follow-up, revealed the importance of a good to excellent Sp (over 95%) over Se for the diagnostic test. Small divergence from perfect Sp extended quickly to disease incidence over-estimation as true prevalence increased and true incidence decreased. Selection and misclassification biases of a low prevalent and incident disease, diagnosed with close to perfect Sp, were minimal, reflecting the importance of choosing a highly specific test to improve unit at risk and case identification. A highly sensitive test to exclude diseased subjects at baseline was of less importance to minimize bias than using a highly specific one at this time point. Of course, the situation would be different in a population with a very high disease prevalence. For most diseases, however, the tendency is to have a large proportion of healthy animals and a small proportion of diseased ones. The range of diseases prevalence investigated in our study (5–20%) would therefore cover most disease scenarios seen in veterinary, and perhaps, human studies.

Near perfect diagnostic test attributes were even more important to obtain a measure of association close to the true risk ratio, according to specific disease characteristics, especially its prevalence. Low prevalent and high incident disease led to minimal bias if disease was diagnosed with high Se and close to perfect Sp. For more prevalent diseases we observed large risk ratio biases towards the null value, even with near perfect diagnosis. This bias also got larger as incidence decreased. For diseases with moderate to high prevalence (20%), the biases could be so important that a study using a test with a Se or Sp < 0.95 would have very little power to identify any measure of association with exposures. Even with prevalence of disease of 5%, a dramatic loss of power is to be expected when imperfect tests are used. Therefore a corollary result of a sub-optimal Sp is that, by causing a bias towards the null, weaker associations (like our RR ~ 1.5) will be more difficult to demonstrate. It would be unnecessary to fight this loss in power by increasing the study sample size in order to get a narrower confidence interval, as the measured association would be biased anyway (Brenner and Savitz, 1990). It was already demonstrated that study power decreases as misclassification increases (Brown and Jiang, 2010). For stronger associations and in the presence of small biases, sample size could be adjusted (Dendukuri et al., 2004; Cheng et al., 2009). But in the presence of larger biased associations towards the null, a weaker, reduced, association would be candidate for further investigation, even if its confidence interval includes 1.0 (Baird et al., 1991).

It is already recognized that misclassification of outcome or exposure during follow-up leads to bias towards null in the estimated associations (Bross, 1954; Copeland et al., 1977; Flegal et al., 1986) as well

as reduced statistical efficiency by loss of power (White, 1986) and confidence intervals of the parameters estimates that are too narrow (Neuhaus, 1999). However this bias towards the null value is strictly true only when misclassification is the same in the two compared groups, i.e. exposure and covariates status do not influence Se and/or Sp (Copeland et al., 1977; Sorahan and Gilthorpe, 1994; Neuhaus, 1999). In this case, we have non-differential misclassification. As shown previously by Copeland et al. (1977), misclassification bias depends primarily on the Sp of the test used and increase with disease rarity, with most of the bias occurring even before the Sp drops below 85%. With Se and Sp as high as 0.90 and 0.96, respectively, RR is already substantially biased (1.5 instead of 2) (Copeland et al., 1977), but when Sp is perfect, bias is absent (Poole, 1985). When disease frequency is low, error in disease diagnosis leads to an increase in false positives which submerge true positives and dilute measures of incidence and association. Bias in RR increases as Se increase and Sp decrease (White, 1986). Exposure misclassification alone can cause serious bias on the RR even if Se or Sp are not lower than 80% (Kristensen, 1992).

When misclassification is differential, i.e. Se and Sp of outcome classification is not equal in each true category of exposure (or Se and Sp of exposure classification is not equal in each true category of outcome), direction of bias for parameter estimates can be in any direction (Dosemeci et al., 1990; Neuhaus, 1999; Chen et al., 2013). In this case, Se and Sp as low as 90% can be sufficient to produce high bias (Kristensen, 1992). Direction of the bias can also be in any direction with dependent misclassification (i.e. the errors in one variable are associated with the errors in an other, Assakul and Proctor, 1967; Greenland, 1989), even if non-differential (Kristensen, 1992). The same is found when the exposure variable is not dichotomous but has multiple levels (Dosemeci et al., 1990; Weinberg et al., 1994). Bias towards the null also requires that selection bias and confounding are absent (Jurek, 2005). There are therefore many situations where bias towards null do not apply. Even when non-differential misclassification is thought to take place, random errors in the observed estimates can lead bias away from the null (Jurek, 2005).

In cohort studies, non-differential misclassification of disease at baseline, i.e. selection bias, especially imperfect Se, can lead to over- or under-estimation of the observed RR (Pekkanen et al., 2006). This bias can be significant for disease with a low true incidence, a high true prevalence, a substantial disease duration (i.e. as long as the interval between first and second test), and a poor test Se. In the presence of misclassification of disease at baseline the observed RR depend on the association between exposure and disease both at baseline and during follow-up (Pekkanen et al., 2006). Therefore to minimize bias, the standard recommendation is to exclude subjects with the outcome at baseline from the cohort based on a highly sensitive test (Pekkanen and Sunyer, 2008). Then during the follow-up period, case identification should use a highly specific test having a high positive predictive value (Brenner and Gefeller, 1993). However Haine et al. (2018) have shown that a more prevalent and incident disease diagnosed with an imperfect Se and/or Sp will give biased measure of association despite attempts to improve its diagnosis.

We have shown here that combined misclassification at baseline and follow-up requires a highly specific test. If a test with high Sp cannot be used, one could use a less efficient test twice at recruitment or for identifying incident cases and with a serial interpretation. The loss in Se of such an approach would cause little bias, compared to the potential gains due to the increased Sp. However, this combined misclassification would also require a highly sensitive test to estimate an association close to the true RR. Unfortunately increasing Sp of a test very often decreases its Se, i.e. a lower probability for diseased individuals to be recognized as diseased. As a results, some classification errors are to be expected leading to biased parameters estimates. If classification errors cannot be avoided during the study design stage, the misclassification bias can be corrected into the analytic stage. For instance, Se and Sp of the test can be incorporated into the modelling strategy (Magder and Hugues, 1997), by performing a probabilistic

sensitivity analysis (Fox et al., 2005), or by including the uncertainty in the estimates with a Bayesian analysis in the form of prior distributions (McInturff et al., 2004). A latent class model (Hui and Walter, 1980) would therefore return the posterior inference on regression parameters and the Se and Sp of both tests. Acknowledgement of these biases and possible corrective measures are important when designing longitudinal studies when gold standard measurement of the outcome might not be readily available, like for bacterial diseases (for example subclinical intramammary infection; Koop et al., 2013), viral diseases (Dotti et al., 2013) or more complex outcome evaluations (e.g. bovine respiratory disease complex; Buczinski et al., 2015). Efforts should be made to improve outcome evaluation but absence or limitation of bias is not always granted in some situation. Haine et al. (2018) demonstrated that for some specific disease incidences and prevalences bias could not be avoided by improving outcome measurements. Using latent class models can help in these cases, as shown by Dufour et al. (2012).

Bias in parameters estimates can be important when considering selection and misclassification biases together in a cohort study. Our results underscore the need for a careful evaluation of the best available options to identify at risk and incident cases according to the expected disease prevalence and incidence of the study.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

DH conducted the simulations, data analysis, results interpretation, and the manuscript writing. ID and SD contributed in interpreting the results and editing the manuscript. DH, ID, and SD contributed to the planning of the study.

FUNDING

This research was financed by the senior author (SD) Natural Sciences and Engineering Research Council of Canada Discovery Grant.

SUPPLEMENTARY FIGURES

Figure S1. Estimated incidence rate as a function of test sensitivity and specificity, a disease prevalence of 5%, and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) when using an imperfect test at baseline (selection bias) or at follow-up (misclassification bias). True incidence rate is found at the upper right corner (i.e. perfect sensitivity and specificity). **Figure S2.** Estimated incidence rate as a function of test sensitivity and specificity, a disease prevalence of 20%, and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) when using an imperfect test at baseline (selection bias) or at follow-up (misclassification bias). True incidence rate is found at the upper right corner (i.e. perfect sensitivity and specificity). **Figure S3.** Estimated risk ratio as a function of test sensitivity and specificity, a disease prevalence of 5%, and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) for an exposure with a true measure of association corresponding to a risk ratio of 3.0 when using an imperfect test at baseline (selection bias) or at follow-up (misclassification bias). True risk ratio is found at the upper right corner (i.e. perfect sensitivity and specificity). **Figure S4.** Estimated risk ratio as a function of test sensitivity

and specificity, a disease prevalence of 20%, and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) for an exposure with a true measure of association corresponding to a risk ratio of 3.0 when using an imperfect test at baseline (selection bias) or at follow-up (misclassification bias). True risk ratio is found at the upper right corner (i.e. perfect sensitivity and specificity). **Figure S5.** Estimated risk ratio as a function of test sensitivity and specificity, disease prevalence (5 or 20%), and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) for an exposure with a true measure of association corresponding to a risk ratio of 1.5 when using an imperfect test both at baseline and follow-up (i.e. total bias). True risk ratio is found at the upper right corner (i.e. perfect sensitivity and specificity). **Figure S6.** Estimated risk ratio as a function of test specificity and disease risk, and for a sensitivity of 95%, when using an imperfect test both at baseline and follow-up. True risk ratio = 1.5. **Figure S7.** Estimated risk ratio as a function of test sensitivity and specificity, a disease prevalence of 5%, and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) for an exposure with a true measure of association corresponding to a risk ratio of 1.5 when using an imperfect test at baseline (selection bias) or at follow-up (misclassification bias). True risk ratio is found at the upper right corner (i.e. perfect sensitivity and specificity). **Figure S8.** Estimated risk ratio as a function of test sensitivity and specificity, a disease prevalence of 20%, and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) for an exposure with a true measure of association corresponding to a risk ratio of 1.5 when using an imperfect test at baseline (selection bias) or at follow-up (misclassification bias). True risk ratio is found at the upper right corner (i.e. perfect sensitivity and specificity).

REFERENCES

Assakul, K. and Proctor, C. H. (1967). Testing independence in two-way contingency tables with data subject to misclassification. *Psychometrika* 32, 67–76. doi:10.1007/bf02289405

Baird, D. D., Weinberg, C. R., and Rowland, A. S. (1991). Reporting errors in time-to-pregnancy data collected with a short questionnaire. *American Journal of Epidemiology* 133, 1282–1290. doi:10.1093/oxfordjournals.aje.a115840

Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics* 33, 414–418. doi:10.1002/0471667196.ess0146.pub2

Brenner, H. and Gefeller, O. (1993). Use of the positive predictive value to correct for disease misclassification in epidemiological studies. *American Journal of Epidemiology* 138, 1007–1015

Brenner, H. and Savitz, D. A. (1990). The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. *American Journal of Epidemiology* 132, 181–192. doi:10.1093/oxfordjournals.aje.a115630

Bross, I. (1954). Misclassification in 2 x 2 tables. *Biometrics* 10, 478–486

Brown, P. and Jiang, H. (2010). Simulation-based power calculations for large cohort studies. *Biometrical Journal* 52, 604–615. doi:10.1002/bimj.200900277

Buczinski, S., Ollivett, T. L., and Dendukuri, N. (2015). Bayesian estimation of the accuracy of the calf respiratory scoring chart and ultrasonography for the diagnosis of bovine respiratory disease in pre-weaned dairy calves. *Preventive Veterinary Medicine* 119, 227–231. doi:https://doi.org/10.1016/j.prevetmed.2015.02.018

Chen, Q., Galfalvy, H., and Duan, N. (2013). Effects of disease misclassification on exposure–disease association. *American Journal of Public Health* 103, e67–e73. doi:10.2105/ajph.2012.300995

Cheng, D., Stamey, J. D., and Branscum, A. J. (2009). Bayesian approach to average power calculations for binary regression models with misclassified outcomes. *Statistics in Medicine* 28, 848–863. doi:10.1002/sim.3505

Clements, D. N., Handel, I. G., Rose, E., Querry, D., Pugh, C. A., Ollier, W. E., et al. (2013). Dogslife: A web-based longitudinal study of Labrador Retriever health in the UK. *BMC Veterinary Research* 9, 13. doi:10.1186/1746-6148-9-13

Copeland, K. T., Checkoway, H., McMichael, A. J., and Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology* 105, 488–495

Courant, R., Robbins, H., and Stewart, I. (1996). *What is Mathematics?: An elementary approach to ideas and methods* (New York: Oxford University Press)

Delgado-Rodriguez, M. and Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health* 58, 635–641. doi:10.1136/jech.2003.008466

Dendukuri, N., Rahme, E., Bélisle, P., and Joseph, L. (2004). Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics* 60, 388–397. doi:10.1111/j.0006-341x.2004.00183.x

Dos Santos Silva, I. (1999). *Cancer Epidemiology: Principles and Methods* (Lyon, France: IARC Scientific Publications). doi:10.1002/sim.759

Dosemeci, M., Wacholder, S., and Lubin, J. H. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology* 132, 746–748. doi:10.1093/oxfordjournals.aje.a115716

Dotti, S., Guadagnini, G., Salvini, F., Razzuoli, E., Ferrari, M., Alborali, G. L., et al. (2013). Time-course of antibody and cell-mediated immune responses to Porcine Reproductive and Respiratory Syndrome virus under field conditions. *Research in Veterinary Science* 94, 510–517. doi:https://doi.org/10.1016/j.rvsc.2012.12.003

Dufour, S., Dohoo, I. R., Barkema, H. W., DesCôteaux, L., DeVries, T. J., Reyher, K. K., et al. (2012). Epidemiology of coagulase-negative staphylococci intramammary infection in dairy cattle and the effect of bacteriological culture misclassification. *Journal of Dairy Science* 95, 3110–3124. doi:10.3168/jds.2011-5164

Flegal, K. M., Brownie, C., and Haas, J. (1986). The effects of exposure misclassification on estimates of relative risk. *American Journal of Epidemiology* 123, 736–751. doi:10.1093/oxfordjournals.aje.a114294

Fox, M. P., Lash, T. L., and Greenland, S. (2005). A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International Journal of Epidemiology* 34, 1370–1376. doi:10.1093/ije/dyi184

Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology* 112, 564–569

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health* 79, 340–349. doi:10.2105/ajph.79.3.340

Haine, D., Dohoo, I., Scholl, D., and Dufour, S. (2018). Diagnosing intramammary infection: Controlling misclassification bias in longitudinal udder health studies. *Preventive Veterinary Medicine* 150, 162–167. doi:10.1016/j.prevetmed.2017.11.010

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* 36, 167–171. doi:10.1002/0471667196.ess0146.pub2

Jurek, A. (2005). Proper interpretation of non-differential misclassification effects: Expectations versus observations. *International Journal of Epidemiology* 34, 680–687. doi:10.1093/ije/dyi060

Koop, G., Collar, C. A., Toft, N., Nielen, M., van Werven, T., Bacon, D., et al. (2013). Risk factors for subclinical intramammary infection in dairy goats in two longitudinal field studies evaluated by Bayesian logistic regression. *Preventive Veterinary Medicine* 108, 304–312. doi:https://doi.org/10.1016/j.prevetmed.2012.11.007

317 Kristensen, P. (1992). Bias from nondifferential but dependent misclassification of exposure and outcome.
318 *Epidemiology* 3, 210–215. doi:10.1097/00001648-199205000-00005
319 Magder, L. S. and Hugues, J. P. (1997). Logistic regression when the outcome is measured with uncertainty.
320 *American Journal of Epidemiology* 146, 195–203
321 McInturff, P., Johnson, W. O., Cowling, D., and Gardner, I. A. (2004). Modelling risk when binary
322 outcomes are subject to error. *Statistics in Medicine* 23, 1095–1109. doi:10.1002/sim.1656
323 Neuhaus, J. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*
324 86, 843–855. doi:10.1093/biomet/86.4.843
325 Pekkanen, J. and Sunyer, J. (2008). Problems in using incidence to analyze risk factors in follow-up studies.
326 *European Journal of Epidemiology* 23, 581–584. doi:10.1007/s10654-008-9280-0
327 Pekkanen, J., Sunyer, J., and Chinn, S. (2006). Nondifferential disease misclassification may bias incidence
328 risk ratios away from the null. *Journal of Clinical Epidemiology* 59, 281–289. doi:10.1016/j.jclinepi.
329 2005.07.013
330 Poole, C. (1985). Exceptions to the rule about nondifferential misclassification. *American Journal of*
331 *Epidemiology* 122, 508
332 Quade, D., Lachenbruch, P. A., Whaley, F. S., McClish, D. K., and Haley, R. W. (1980). Effects of
333 misclassifications on statistical inferences in epidemiology. *American Journal of Epidemiology* 111,
334 503–515
335 R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for
336 Statistical Computing, Vienna, Austria
337 Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American*
338 *Journal of Epidemiology* 107, 71–76
339 Rothman, K. J., Lash, T. L., and Greenland, S. (2012). *Modern Epidemiology* (Lippincott Williams &
340 Wilkins)
341 Sorahan, T. and Gilthorpe, M. S. (1994). Non-differential misclassification of exposure always leads to an
342 underestimate of risk: An incorrect conclusion. *Occupational and Environmental Medicine* 51, 839–840.
343 doi:10.1136/oem.51.12.839
344 Tobias, T., Klinkenberg, D., Bouma, A., van den Broek, J., Daemen, A., Wagenaar, J., et al. (2014). A
345 cohort study on Actinobacillus pleuropneumoniae colonisation in suckling piglets. *Preventive Veterinary*
346 *Medicine* 114, 223–230. doi:10.1016/j.prevetmed.2014.02.008
347 Weinberg, C. A., Umbach, D. M., and Greenland, S. (1994). When will nondifferential misclassification
348 of an exposure preserve the direction of a trend? *American Journal of Epidemiology* 140, 565–571.
349 doi:10.1093/oxfordjournals.aje.a117283
350 White, E. (1986). The effect of misclassification of disease status in follow-up studies: Implications for
351 selecting disease classification criteria. *American Journal of Epidemiology* 124, 816–825. doi:10.1093/
352 oxfordjournals.aje.a114458

FIGURE CAPTIONS

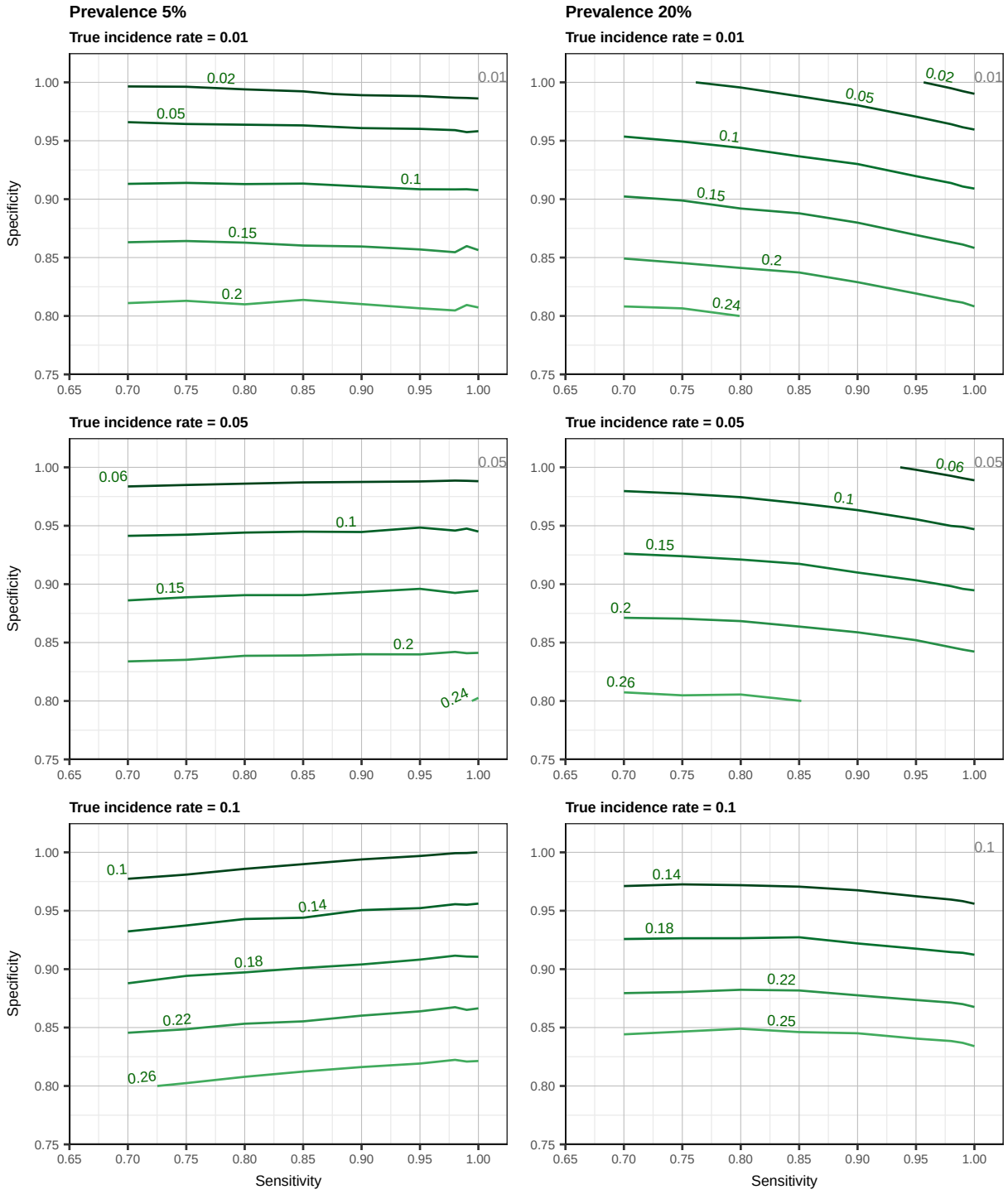

Figure 1. Estimated incidence rate (in cases/animal-time unit) as a function of test sensitivity and specificity, disease prevalence (5 or 20%), and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) when using an imperfect test both at baseline and follow-up (i.e. total bias). True incidence rate is found at the upper right corner (i.e. perfect sensitivity and specificity).

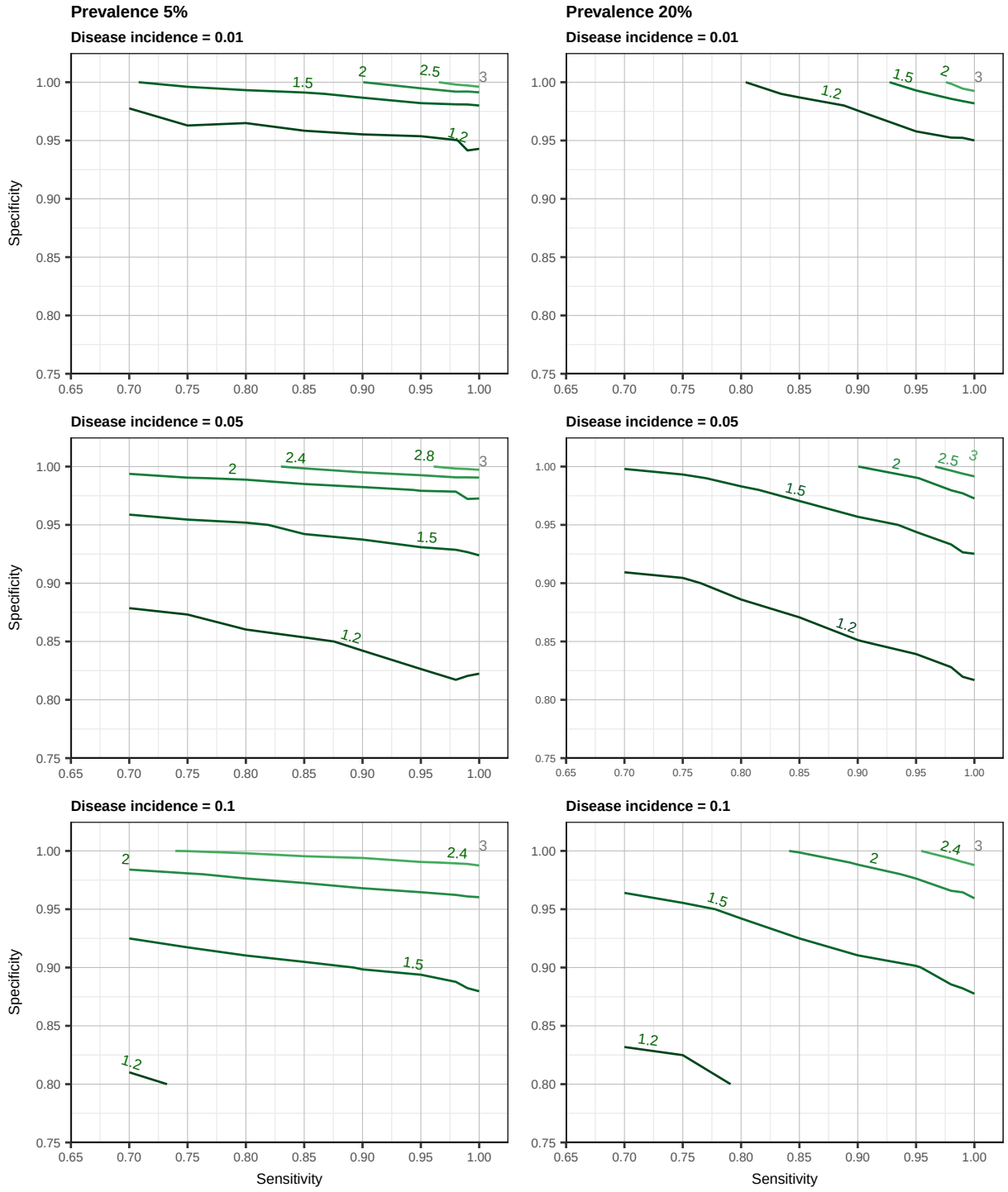

Figure 2. Estimated risk ratio as a function of test sensitivity and specificity, disease prevalence (5 or 20%), and true disease incidence (0.01, 0.05, 0.1 case/animal-time unit) for an exposure with a true measure of association corresponding to a risk ratio of 3.0 when using an imperfect test both at baseline and follow-up (i.e. total bias). True risk ratio is found at the upper right corner (i.e. perfect sensitivity and specificity).

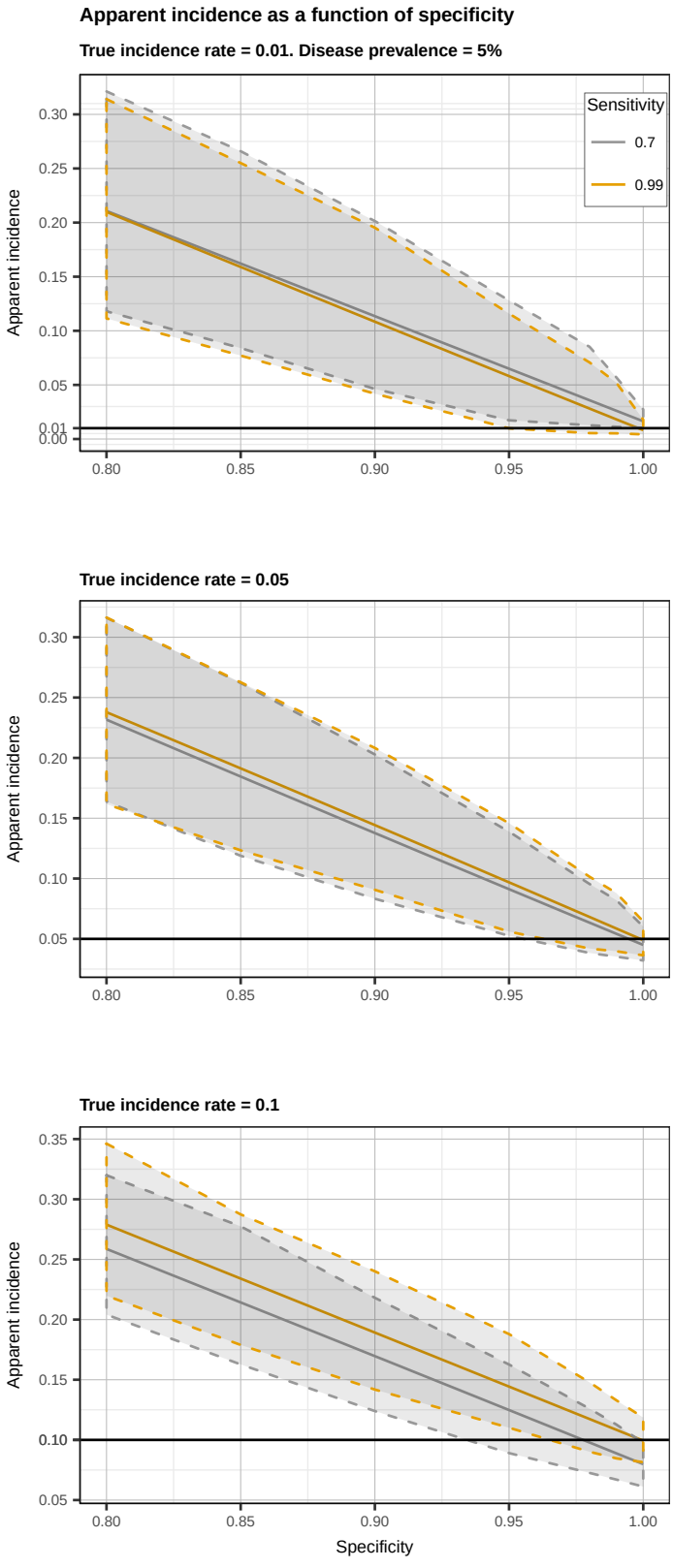

Figure 3. Apparent incidence resulting from total bias, as a function of specificity. Disease prevalence = 5%. Solid line: median value; dotted lines: first and third quartiles.

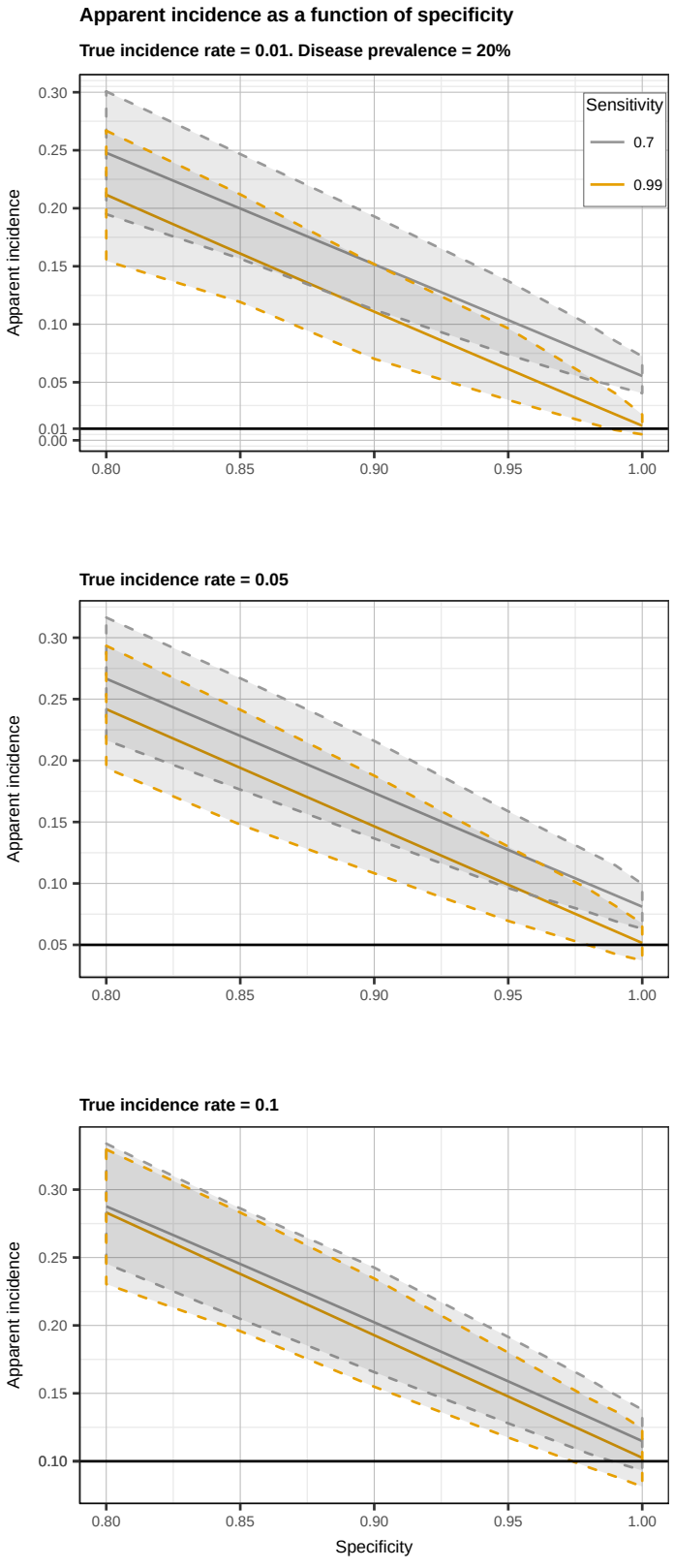

Figure 4. Apparent incidence resulting from total bias, as a function of specificity. Disease prevalence = 20%. Solid line: median value; dotted lines: first and third quartiles.

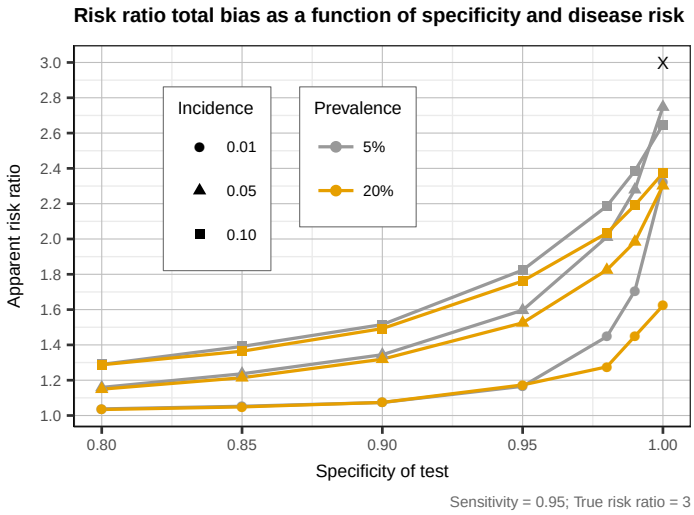

Figure 5. Estimated risk ratio as a function of test specificity and disease risk, and for a sensitivity of 95%, when using an imperfect test both at baseline and follow-up. True risk ratio = 3.0.