

RESEARCH ARTICLE

# Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness

Tom De Smedt<sup>1</sup>, Elizabeth Merrall<sup>2</sup>, Denis Macina<sup>3</sup>, Silvia Perez-Vilar<sup>4,5</sup>, Nick Andrews<sup>6</sup>, Kaatje Bollaerts<sup>1\*</sup>

**1** P95 Epidemiology and Pharmacovigilance, Leuven, Belgium, **2** GSK Vaccines, Amsterdam, The Netherlands, **3** Sanofi Pasteur, Lyon, France, **4** FISABIO-Public Health, Valencia, Spain, **5** Erasmus University Medical Center, Rotterdam, The Netherlands, **6** Statistics, Modelling, and Economics Department, Public Health England, Colindale, London, United Kingdom

\* [Kaatje.Bollaerts@p-95.com](mailto:Kaatje.Bollaerts@p-95.com)



**OPEN ACCESS**

**Citation:** De Smedt T, Merrall E, Macina D, Perez-Vilar S, Andrews N, Bollaerts K (2018) Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. *PLoS ONE* 13(6): e0199180. <https://doi.org/10.1371/journal.pone.0199180>

**Editor:** Daniela Flavia Hozbor, Universidad Nacional de la Plata, ARGENTINA

**Received:** November 22, 2017

**Accepted:** June 1, 2018

**Published:** June 15, 2018

**Copyright:** This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

**Data Availability Statement:** All relevant data are available in the paper. The web-application can be accessed at <http://apps.p-95.com/VEMisclassification/>.

**Funding:** This research was funded by the Innovative Medicines Initiative (IMI) Joint Undertaking through the ADVANCE project [N° 115557]. The IMI is a joint initiative (public-private partnership) of the European Commission and the European Federation of Pharmaceutical Industries and Associations (EFPIA) to improve the

## Abstract

### Background

Studies of vaccine effectiveness (VE) rely on accurate identification of vaccination and cases of vaccine-preventable disease. In practice, diagnostic tests, clinical case definitions and vaccination records often present inaccuracies, leading to biased VE estimates. Previous studies investigated the impact of non-differential disease misclassification on VE estimation.

### Methods

We explored, through simulation, the impact of non-differential and differential disease- and exposure misclassification when estimating VE using cohort, case-control, test-negative case-control and case-cohort designs. The impact of misclassification on the estimated VE is demonstrated for VE studies on childhood seasonal influenza and pertussis vaccination. We additionally developed a web-application graphically presenting bias for user-selected parameters.

### Results

Depending on the scenario, the misclassification parameters had differing impacts. Decreased exposure specificity had greatest impact for influenza VE estimation when vaccination coverage was low. Decreased exposure sensitivity had greatest impact for pertussis VE estimation for which high vaccination coverage is typically achieved. The impact of the exposure misclassification parameters was found to be more noticeable than that of the disease misclassification parameters. When misclassification is limited, all study designs perform equally. In case of substantial (differential) disease misclassification, the test-negative design performs worse.

competitive situation of the European Union in the field of pharmaceutical research. The IMI provided support in the form of salaries for TDS and KB, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The work by EM and DM was covered by EFPIA's in-kind contribution to IMI projects. SPV and NA did not receive any financial compensation for their contribution to this research. The specific roles of these authors are articulated in the 'author contributions' section.

**Competing interests:** At the time of the research, EM was employed by GSK and DM by Sanofi Pasteur. Both companies develop vaccines and support the IMI ADVANCE project. KB received consulting fees from vaccine producing companies (GSK, SPMSD, Pfizer, Takeda) not related to the research presented here. TDS received consulting fees from Pfizer and Takeda not related to this work. At the time of the research, SPV was employed by FISABIO and Erasmus MC. The authors have declared that no competing interests exist.

## Conclusions

Misclassification can lead to significant bias in VE estimates and its impact strongly depends on the scenario. We developed a web-application for assessing the potential (joint) impact of possibly differential disease- and exposure misclassification that can be modified by users to their own study scenario. Our results and the simulation tool may be used to guide better design, conduct and interpretation of future VE studies.

## Introduction

Vaccine effectiveness (VE) is defined as a measure of protection among vaccinated persons attributable to a vaccine administered under field conditions to a given population, which is different from vaccine efficacy being defined as the effect of vaccination among vaccinated persons as measured in pre-licensure clinical trials with vaccination allocated under optimal conditions [1]. Whilst aggregated data may be used for assessment of impact and uptake, individual level data are usually required to estimate VE. Such data may be available nationally, regionally or in health systems which are nationally representative.

When studying VE, it is essential to accurately identify cases of the vaccine preventable disease and the vaccination status (e.g. defined as 1 dose vs none, 2 doses vs 1 dose, or completely vs partially vaccinated, depending on the research question of interest). Indeed, assuming misclassification is non-differential and independent of other errors, both disease and exposure misclassifications tend to bias the VE estimates toward the null [2]. Disease and exposure statuses may reciprocally affect each other's ascertainment (i.e. differential misclassification) and lead to biased estimates in either direction [3]. For example, differential disease misclassification might arise from differences in healthcare seeking behavior, with subjects more likely to seek care being more likely vaccinated and also being more likely correctly diagnosed as diseased. Laboratory confirmation is desirable when assessing VE [4]. However, laboratory test results are not always available or perfectly accurate and, especially in health care database-based analyses, case definitions often rely on clinical criteria, potentially resulting in disease misclassification. Different sources of disease misclassification exist and they might be broadly categorized as under-ascertainment (individuals that do not seek healthcare) and underreporting (individuals that do seek healthcare, but whose health event is not accurately captured due to various reasons) [5]. Likewise, the vaccination exposure information might be subject to coding entry error or omissions potentially biasing estimates of VE as well [6].

Concerns regarding disease and exposure misclassifications are particularly relevant when conducting epidemiological studies using health care databases [7]. Nonetheless, and despite concerns on data validity, sample representativity and the limited ability to control for confounding, there is a strong interest in using large health care databases to study vaccine use and the outcomes of vaccination by projects such as the Vaccine Safety Datalink [8], Post-Licensure Rapid Immunization Safety Monitoring programme [9] and ADVANCE (<http://www.advance-vaccines.eu/>). Indeed, the size of observational databases allows for the study of rare events and, as they are embedded within clinical practice, they offer the potential to study real-world vaccine effects relatively efficiently from both cost and time perspectives.

When conducting VE studies it is important to quantify the potential impact of misclassification on the VE estimates in order to assess study feasibility, optimize study design and possibly, the need to correct for misclassification. In earlier work, the impact of non-differential disease misclassification on influenza VE has been quantified for cohort, case-control and test-

negative designs based on mathematical derivations [10] and using simulation studies [10, 11]. We extended the simulation study by Orenstein [10] and Jackson [10, 11], to account for both disease- and exposure misclassification and allow for both differential and non-differential misclassification. Furthermore, as we show that the impact of misclassification on the estimated VE depends both on the epidemiology of the vaccine preventable disease and the expected vaccination coverage, we developed a web-application allowing to run simulations with user-defined parameters. We illustrate the impact of misclassification on VE estimates using two examples with clearly different disease attack rates and expected vaccination coverage; a) childhood pertussis and b) pediatric seasonal influenza VE estimations.

This work was carried out under the auspices of the ‘Accelerated development of vaccine benefit-risk collaboration in Europe’ (ADVANCE) project, launched in 2013, funded by the Innovative Medicines Initiative (IMI). The aim of ADVANCE is to help health professionals, regulatory agencies, public health institutions, vaccine manufacturers, and the general public make well-informed and timely decisions on benefits and risks of marketed vaccines by establishing a framework and toolbox to enable rapid delivery of reliable data on vaccine benefits and risks.

## Methods

In this section, we first present analytical derivations illustrating the impact of misclassification on VE estimates at population level—hence ignoring estimation error—when considering misclassification in its simplest form, being single source non-differential misclassification. Although estimation error is ignored, such analytical derivations provide meaningful insights. However, the derivations become tedious in situations where misclassification is more complex, especially when considering the joint impact of disease and exposure misclassification. Therefore, we also assess through simulation the impact of differential and non-differential disease and exposure misclassification when estimating VE using cohort, case-control, test-negative case-control and case-cohort (screening method) designs. These designs are used to estimate VE, with the classical cohort and case-control designs being probably the most commonly used ones [12]. The test-negative case-control design is popular for estimating VE of vaccines for influenza and rotavirus [13]. In the test-negative design, the study population are patients who are seeking medical care for a defined clinical condition (e.g. acute respiratory illness) and are tested for a specific viral infection (e.g. influenza). Then, patients testing positive are the cases and patients testing negative are the controls. Finally, the case-cohort or screening method uses data on the exposure prevalence in cases and compares this to the exposure prevalence from an external coverage cohort, from which the cases originate [14].

## Notation

First, let  $\pi_{VPD,0}$  be the unobserved ‘true’ risk of disease due to the pathogen targeted by the vaccine (vaccine preventable disease, VPD) in unvaccinated subjects,  $\pi_{Other}$  the corresponding risk of similar disease due to other pathogens than those targeted by the vaccine, and let  $\gamma$  be the ‘true’ vaccination coverage. Vaccination affects the VPD risk, with the risk among the vaccinated  $\pi_{VPD,1} = (1 - VE)\pi_{VPD,0}$ , but does not affect the other disease risk. Furthermore, let  $p_0$  be the observed disease prevalence among the subjects indicated as unvaccinated and  $p_1$  the observed prevalence among the subjects indicated as vaccinated. Finally, let  $SE_d$  be the disease sensitivity (probability of being indicated as diseased if truly diseased) and  $SP_d$  the disease specificity (probability of being indicated as not diseased if truly not diseased) of the case definition. Similarly let  $SE_e$  be the exposure sensitivity (probability of being indicated as exposed if truly exposed) and  $SP_e$  the exposure specificity (probability of being indicated as unexposed if

truly unexposed) of the exposure ascertainment definition. In the case of differential misclassification, the disease misclassification parameters depend on exposure status and vice versa, yielding four disease misclassification parameters;  $SE_{d,E=0}$ ,  $SE_{d,E=1}$ ,  $SP_{d,E=0}$ ,  $SP_{d,E=1}$  (with  $E = 0$  indicating unvaccinated subjects and  $E = 1$  vaccinated subjects) and four exposure misclassification parameters;  $SE_{e,D=0}$ ,  $SE_{e,D=1}$ ,  $SP_{e,D=0}$ ,  $SP_{e,D=1}$  (with  $D = 0$  indicating not diseased subjects and  $D = 1$  diseased subjects).

### Impact of misclassification at population-level

**Non-differential disease misclassification.** Given the simplifying assumptions of no exposure misclassification and no co-infection between the VPD and the similar disease due to other pathogens, the observed disease risk among the unvaccinated is the sum of the probability of having the VPD and being correctly indicated as such (true positive for disease) and the probability of having the non-VPD and being incorrectly indicated as having the VPD (false positive for disease) or

$$p_0 = SE_d \pi_{VPD,0} + (1 - SP_d)\pi_{Other}. \tag{1}$$

Similarly, for the vaccinated, the observed disease risk equals

$$p_1 = SE_d \pi_{VPD,1} + (1 - SP_d)\pi_{Other}, \tag{2}$$

with  $\pi_{VPD,1} = (1 - VE)\pi_{VPD,0}$ .

In line with Orenstein [10] and analogous to the statistical definition of bias, we define the population-level bias as the difference in VE for a population with and without misclassification or

$$\Delta = \left(1 - \frac{p_1}{p_0}\right) - \left(1 - \frac{\pi_{VPD,1}}{\pi_{VPD,0}}\right) = \frac{\pi_{VPD,1}}{\pi_{VPD,0}} - \frac{SE_d \pi_{VPD,1} + (1 - SP_d)\pi_{Other}}{SE_d \pi_{VPD,0} + (1 - SP_d)\pi_{Other}}. \tag{3}$$

This expression can be rewritten as

$$\Delta = \frac{(\pi_{VPD,1} - \pi_{VPD,0})(1 - SP_d)\pi_{Other}}{\pi_{VPD,0}(SE_d \pi_{VPD,0} + (1 - SP_d)\pi_{Other})}, \tag{4}$$

showing that the bias equals zero if the disease specificity equals one, and this irrespective of the disease sensitivity.

Now, solving (1) for  $\pi_{VPD,0}$  and (2) for  $\pi_{VPD,1}$ , we have

$$\pi_{VPD,0} = (p_0 - (1 - SP_d)\pi_{Other})/SE_d. \tag{5}$$

$$\pi_{VPD,1} = (p_1 - (1 - SP_d)\pi_{Other})/SE_d, \tag{6}$$

based on which, and given accurate estimates of disease misclassification parameters, an estimate of the ‘true’ VE corrected for disease misclassification can be obtained as

$$VE_\pi = 1 - \frac{p_1 - (1 - SP_d)\pi_{Other}}{p_0 - (1 - SP_d)\pi_{Other}}. \tag{7}$$

Interestingly, the correction equation requires an estimate of disease specificity but not of disease sensitivity. Obviously, the latter only holds if the disease misclassification is non-differential by vaccination status.

**Non-differential exposure misclassification.** Given the simplifying assumption of no disease misclassification the disease prevalence among subjects indicated as unvaccinated is

the sum of the probability of having the VPD and being incorrectly indicated as unvaccinated (false negative for vaccination), and the probability of having the VPD and being correctly indicated as unvaccinated (true negative for vaccination) or

$$p_0 = (1 - SE_e) \gamma \pi_{VPD,1} + SP_e(1 - \gamma)\pi_{VPD,0}, \tag{8}$$

with true vaccination coverage  $\gamma$ . Similarly, the true positives and false positives for vaccination determine the disease risk among the subjects indicated as vaccinated or

$$p_1 = SE_e \gamma \pi_{VPD,1} + (1 - SP_e)(1 - \gamma) \pi_{VPD,0}. \tag{9}$$

The population-level bias due to exposure misclassification is now defined as

$$\Delta = \left(1 - \frac{p_1}{p_0}\right) - \left(1 - \frac{\pi_{VPD,1}}{\pi_{VPD,0}}\right) = \frac{\pi_{VPD,1}}{\pi_{VPD,0}} - \frac{SE_e \gamma \pi_{VPD,1} + (1 - SP_e)(1 - \gamma) \pi_{VPD,0}}{(1 - SE_e) \gamma \pi_{VPD,1} + SP_e(1 - \gamma)\pi_{VPD,0}}. \tag{10}$$

This expression shows that the impact of sensitivity will be largest when coverage is high whereas the impact of specificity will be largest when coverage is low.

Solving (8) and (9) for  $\pi_{VPD,0}$  and for  $\pi_{VPD,1}$ , we obtain

$$\pi_{VPD,0} = (p_0 SE_e - p_1(1 - SE_e))/((1 - \gamma)(SE_e + SP_e - 1)), \tag{11}$$

$$\pi_{VPD,1} = (p_1 SP_e - p_0(1 - SP_e))/(\gamma(SP_e + SE_e - 1)). \tag{12}$$

Then, an expression of the ‘true’ VE corrected for exposure misclassification corresponds to

$$VE_\pi = 1 - \left(\frac{1 - \gamma}{\gamma}\right) \frac{p_1 SP_e - p_0(1 - SP_e)}{p_0 SE_e - p_1(1 - SE_e)}. \tag{13}$$

This correction equation depends—next to the observed disease risks—on both exposure sensitivity and specificity as well as on the ‘true’ vaccination coverage.

### Simulation tool

Similar to Jackson [11], we simulate populations at risk for two outcomes; the VPD and a comparable outcome due to infection with one or more pathogen(s) not targeted by the respective vaccination. We assume that a number of subjects are vaccinated with coverage  $\gamma$ . Unvaccinated subjects could develop the VPD (only once) with a risk equal to  $\pi_{VPD,0}$  and the health outcome due to infection with other pathogens (only once) with a risk equal to  $\pi_{other}$ . For vaccinated subjects, the risk of developing the VPD is reduced to  $\pi_{VPD,1} = (1 - VE)\pi_{VPD,0}$ , whereas the risk due to other pathogens is unaffected by vaccination. We furthermore assume that the risks of developing both outcomes are independent. After having allocated the ‘true’ disease- and exposure status, we randomly allow these events to be misclassified. In particular, for the disease events, diseased cases are misclassified as not diseased with a probability of  $1 - SE_d$  and not diseased cases are misclassified as diseased with a probability of  $1 - SP_d$ . The same holds for the exposure events, but using the exposure sensitivity  $SE_e$  and specificity  $SP_e$  parameters to simulate misclassification. In the case of differential misclassification, the disease misclassification parameters depend on exposure status and vice versa, yielding eight misclassification parameters in total; four disease misclassification parameters;  $SE_{d,E=0}$ ,  $SE_{d,E=1}$ ,  $SP_{d,E=0}$ ,  $SP_{d,E=1}$  and four exposure misclassification parameters;  $SE_{e,D=0}$ ,  $SE_{e,D=1}$ ,  $SP_{e,D=0}$ ,  $SP_{e,D=1}$ .

Then, for a given parameter setting, a large number of simulated populations ( $k = 1, 2, \dots, K$ ) of a predefined population size  $N$  are generated. Based on the observed exposure and disease statuses in each population  $k$ , VE is estimated using the cohort, case-control, test-negative

**Table 1. Estimation of vaccine effectiveness (VE) for the cohort, case-control, test-negative case-control and case-coverage (screening method) design.**

	Cohort	Case-control	Test-negative case-control	Screening method
For each simulated population	We calculate the VPD risk in the vaccinated vs in the unvaccinated.	We identify cases of VPD and sample controls from the full population at risk (case-cohort sampling); and for these two groups compare the odds of exposure as an odds ratio. We used case-cohort sampling as it was recommended in [10].	Here, the cases are the outcome events due to the VPD pathogen (test-positives) and the controls are the outcome events due to other pathogens (test-negatives).	We use only the exposure statuses of the observed cases and compare the odds of exposure in these cases with the odds of exposure in the external coverage cohort.
Estimate VE as	$\widehat{VE}_{Co} = 1 - \widehat{RR}_{Co}$ $= 1 - \frac{\widehat{p}_v}{\widehat{p}_u},$ with $\widehat{RR}_{Co}$ the estimated ratio of the VPD risk in the vaccinated vs. unvaccinated; and estimated risks $\widehat{p}_v$ and $\widehat{p}_u$ based on observed proportions of VPD in the vaccinated and unvaccinated respectively.	$\widehat{VE}_{CC} = 1 - \widehat{OR}_{CC}$ $= 1 - \frac{\widehat{p}_d/(1 - \widehat{p}_d)}{\widehat{p}_n/(1 - \widehat{p}_n)},$ with $\widehat{OR}_{CC}$ the estimated ratio of odds of exposure in cases vs. controls, which is equivalent to the odds of VPD in the vaccinated versus unvaccinated; and $\widehat{p}_d$ and $\widehat{p}_n$ being the observed proportions of exposure in the cases and controls respectively.	$\widehat{VE}_{TN} = 1 - \widehat{OR}_{TN}$ $= 1 - \frac{\widehat{p}_{ip}/(1 - \widehat{p}_{ip})}{\widehat{p}_{in}/(1 - \widehat{p}_{in})},$ with $\widehat{OR}_{TN}$ the estimated ratio of the odds of exposure in the cases versus controls; $\widehat{p}_{ip}$ and $\widehat{p}_{in}$ observed proportions of exposure in test-positive and test-negative individuals respectively.	$\widehat{VE}_{SCREEN} = 1 - \widehat{OR}_{SCREEN}$ $= 1 - \frac{\widehat{p}_d/(1 - \widehat{p}_d)}{\widehat{X}/(1 - \widehat{X})},$ with $\widehat{OR}_{SCREEN}$ being the estimated ratio of the odds of exposure in the cases vs the odds of exposure in the external coverage cohort; $\widehat{p}_d$ is as defined for the case-control design and $\widehat{X}$ an estimate of the vaccine coverage for the external coverage cohort. For the simulation model, $\widehat{X}$ is estimated as the proportion of individuals with observed exposures, assuming same levels of misclassification in the external coverage cohort as in the cases.

<https://doi.org/10.1371/journal.pone.0199180.t001>

case-control and case-coverage designs, using case-cohort sampling as recommended in [10, 11] for the case-control designs (Table 1). Then, these estimates are compared with the true VE used to generate the simulated populations. The biases are compared graphically.

The simulation model is developed using R 3.3.1 [15]. To allow modifying the simulations for other parameter settings/diseases while maximizing user-friendliness, we have encapsulated the source code of the simulation model in a web application created using the Shiny package [16]. Through the web application, the user can set all the necessary input parameters and the output files can be downloaded. The application can be found at the ADVANCE website (<http://www.advance-vaccines.eu/>) or at <http://apps.p-95.com/VEMisclassification/>.

### Scenarios

**General settings.** In this paper, we present two specific vaccination scenarios, pediatric seasonal influenza and childhood pertussis vaccination. For each subsequent simulation scenario, we set  $K = 1000$  and  $N = 50\,000$  whereas VE, vaccination coverage and the respective attack rates depend on the specific scenarios detailed below. We vary one-by-one the disease- or exposure misclassification from  $\{0.50, 0.60, \dots\}$  while fixing the remaining misclassification parameters to 1.

**Pediatric seasonal influenza.** For consistency with Orenstein [10] and Jackson [11], we assumed a 1-dose VE of 70%, an attack rate (AR) of influenza in the unvaccinated of 15% and an AR of influenza-like illness not caused by influenza of 30%. The pediatric seasonal influenza vaccination coverage was assumed to be 10%, in line with the coverage rates reported for the majority of European countries [17].

**Pertussis primary series.** We assumed a VE of 80%, derived as a conservative value from a Cochrane systematic review of vaccine efficacy estimates obtained in random clinical trials, which found the efficacy of acellular pertussis vaccines in pediatric primary series to range between 71% and 85% for a follow-up period ranging from 17 to 22 months after vaccination

[18]. We furthermore assumed that the AR of pertussis in the unvaccinated was 15% [19] and the AR of the non-vaccine preventable pathogens was 10.5% [20]. For the vaccination coverage, we assumed a value of 95%, which reflects a coverage rate commonly reported for the pediatric primary series in high-income countries[21].

## Results

### Pediatric seasonal influenza

In the seasonal influenza scenario and assuming non-differential misclassification (Fig 1, left), the exposure specificity had the largest impact when fixing the remaining parameters to 1 followed by disease specificity and this across all designs. Indeed, the VE was most strongly underestimated when lowering the exposure specificity from 1 to 0.5. The underestimation in VE was still pronounced but less when lowering the disease specificity. Lowering the exposure sensitivity had a negligible impact on the VE whereas lowering the disease sensitivity had no impact when the remaining parameters were fixed to 1.

In case of differential exposure misclassification (Fig 1, middle), the bias could go in either direction, with the estimated VE showing very large deviations from the true VE. Across all designs, the exposure specificity for the diseased had the strongest impact among all four exposure misclassification parameters when fixing the remaining parameters to 1 and biases the VE estimates downwards. Also the exposure sensitivity for the undiseased yields a downwards bias. Lowering the exposure sensitivity for the diseased and the exposure specificity for the undiseased both show a slightly upwards bias.

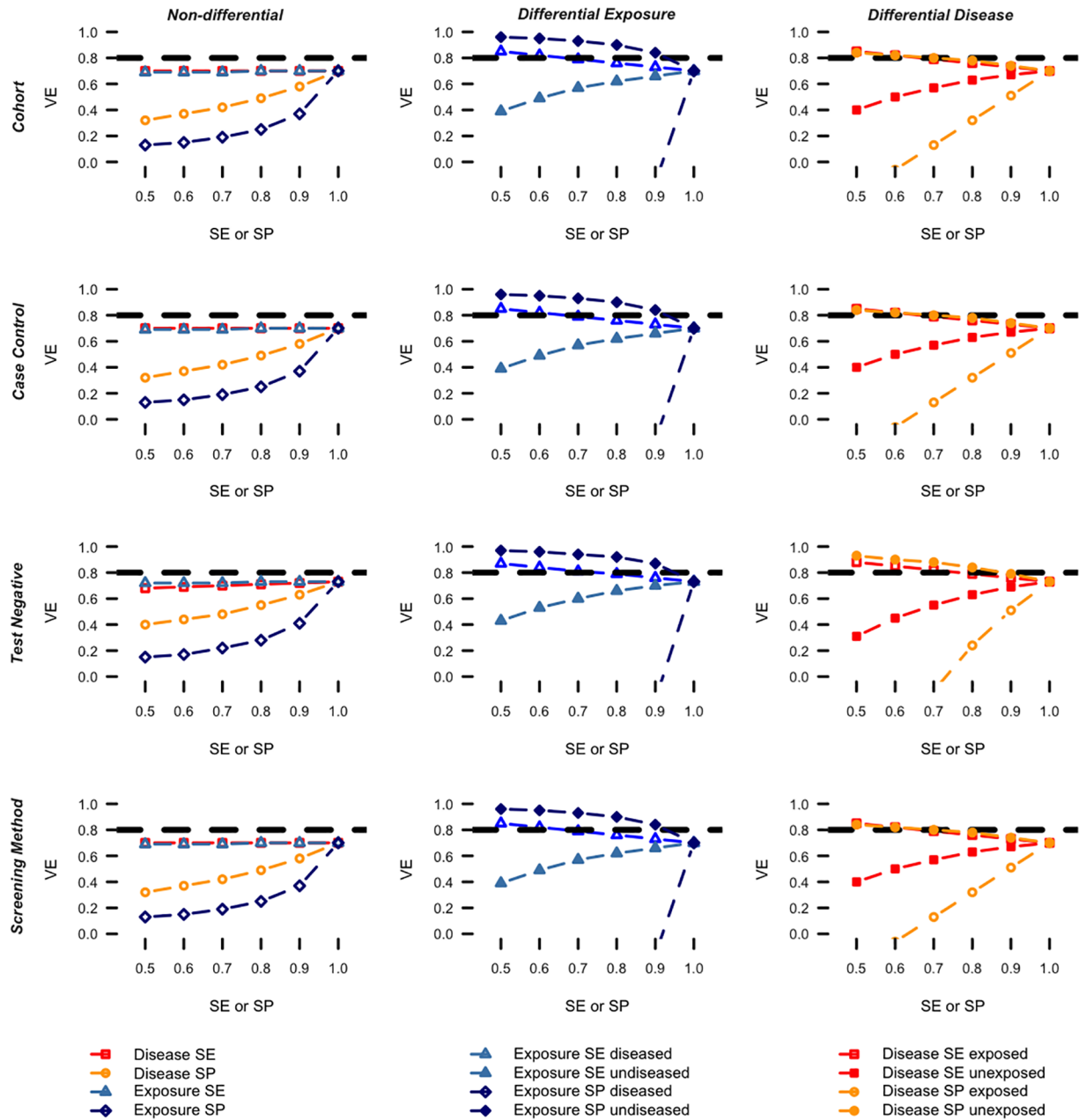
In case of differential disease misclassification (Fig 1, right), the bias could go in either direction as well. Across all designs, the disease specificity for the exposed had the largest (downwards biasing) impact among the four disease misclassification parameters when fixing the remaining parameters to 1 followed by the disease sensitivity in the unexposed. The disease sensitivity for the exposed and the disease specificity for the unexposed are both associated with a slightly upwards bias. The test negative design performs worse than the other designs, particularly for low levels of disease specificity in the exposed.

### Pertussis primary series

In the pertussis scenario and assuming non-differential misclassification (Fig 2, left), the exposure sensitivity had the largest impact when fixing the remaining parameters to 1 followed by disease specificity. In case of differential exposure misclassification (Fig 2, middle), the exposure sensitivity for the un-diseased had the strongest impact among all four exposure misclassification parameters and biased the VE estimates downwards. Finally, in case of differential disease misclassification (Fig 2, right), the disease specificity for the exposed had the largest impact among the four disease misclassification parameters. The impact of the misclassification parameters was comparable across designs. As with pediatric influenza, the bias due to differential misclassification could go in either direction and lead to very large deviations from the true VE. Again, misclassification more strongly affects VE estimates from test-negative designs.

## Discussion

The development of the simulation tool has presented an opportunity to explore the interplay of disease- and exposure misclassification in VE estimations from different study designs. In this study, we explored the single impact of non-differential and differential disease- and exposure misclassification on childhood seasonal influenza and pertussis VE

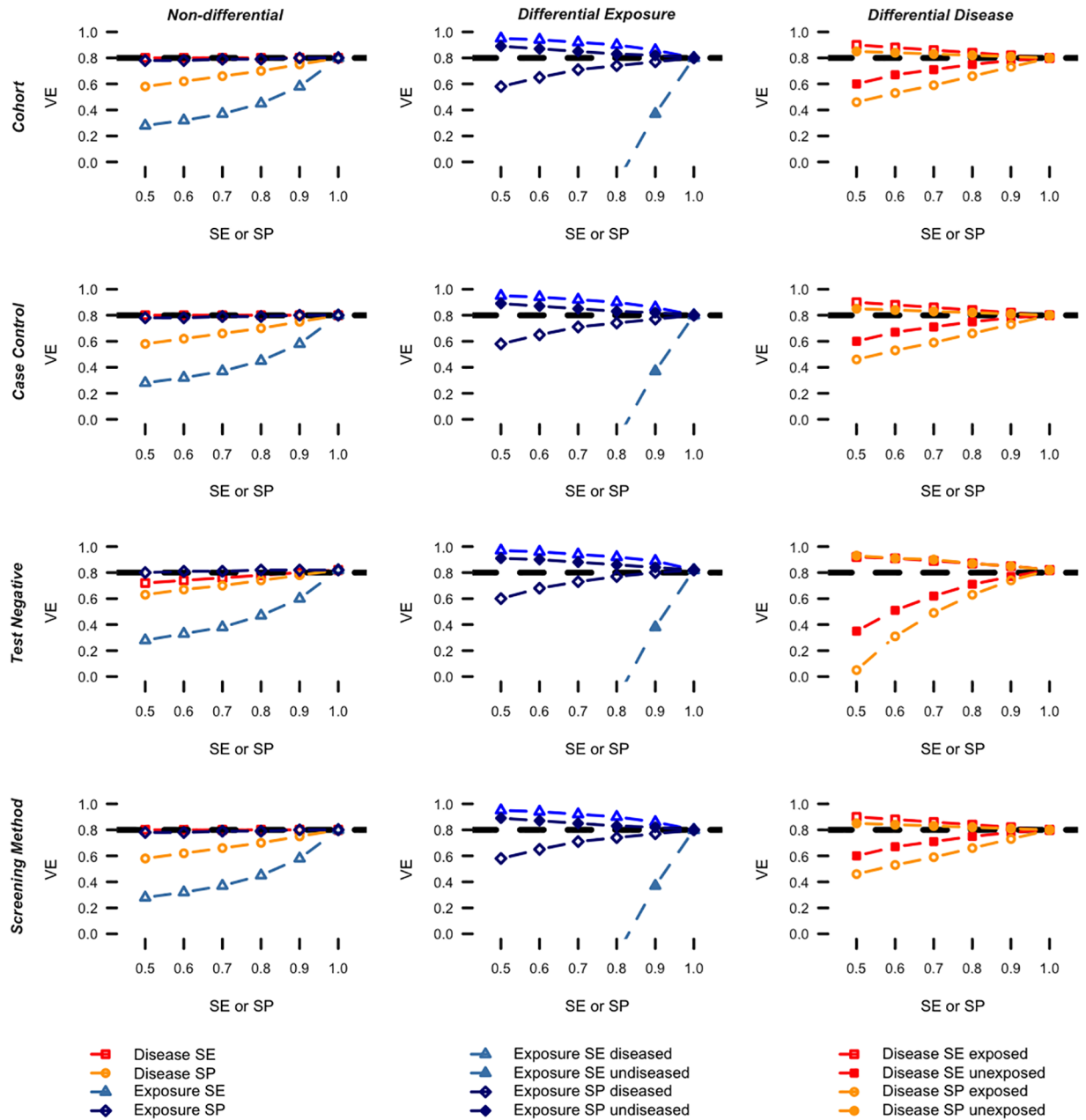


**Fig 1. Influenza scenario: Vaccine effectiveness by design for varying levels of exposure- and disease misclassification while fixing the remaining parameters to 1.** The dashed horizontal lines indicate the true VE used to simulate the data.

<https://doi.org/10.1371/journal.pone.0199180.g001>

estimation. Depending on the scenario, the misclassification parameters had differing impacts. Decreased exposure specificity (poorer identification of non-vaccinees) had greatest impact for influenza VE estimation. Conversely decreased exposure sensitivity (poorer identification of vaccinees) had greatest impact for Pertussis VE estimation. These different impacts correspond to the respectively low and high vaccine coverage in the two scenarios, which is also supported by the analytical derivation (10) in Section 2.2. Similar observations were made regarding the impact of the exposure prevalence on the predictive values of the exposure assessment. Indeed, in low prevalence settings, the exposure specificity has the





**Fig 2. Pertussis scenario: Vaccine effectiveness by design for varying levels of exposure- and disease misclassification while fixing the remaining parameters to 1.** The dashed horizontal lines indicate the true VE used to simulate the data.

<https://doi.org/10.1371/journal.pone.0199180.g002>

greatest impact with the lower the specificity the lower the positive predictive value. Conversely, in high prevalence settings, the exposure sensitivity has the greatest impact with the lower the sensitivity the lower the negative predictive value. Finally, it is interesting to note that, for the influenza and pertussis scenarios investigated, we found exposure misclassification to have a larger impact compared to disease misclassification whereas previous research focused on disease misclassification only.

The impact of the misclassification parameters was found to be more noticeable than that of the different study designs, with the different study designs performing similarly when

misclassification is limited. Jackson [11] found earlier that VE estimates from test-negative case-control designs are more biased than those from classical cohort and case-control designs in case of substantial non-differential disease misclassification. We were able to replicate these results, and also found that the test-negative design performs particularly worse in case of substantial differential disease misclassification, with strong downward biases for low levels of disease specificity in the exposed. The worse performance of the test-negative design can be intuitively explained by comparing the case-control and test-negative design. In case-control designs, the false positives originate from the entire population of subjects free of the VPD with the false positive risk equal to the product of non-VPD risk and 1 minus the disease specificity. On the other hand, in test-negative designs, the false positives originate from the population of test-negatives with the false positive risk equal to 1 minus the disease specificity. Hence, the relative number of perturbations due to falsely classifying controls as cases is much smaller for the classical case-control design compared to the test-negative design.

Although the test-negative design is more sensitive to disease misclassification compared to other designs, its performance remains good when misclassification is limited. Next to misclassification, other sources of bias such as confounding and selection bias should be considered when selecting an appropriate study design. For instance, observational studies on influenza VE might be strongly confounded by differences in healthcare seeking behavior between vaccinated and unvaccinated persons, therefore the test-negative design might still be the appropriate choice in this case [22].

The dependence of the impact of misclassification on the scenario urged us to develop a user-friendly simulation tool that can be modified by users to their own study scenario. The tool allows users to assess the single and joint impact of both differential and non-differential disease- and exposure misclassification on VE estimates from cohort, case-control, test-negative case-control and case-coverage studies. The simulation tool can be accessed through the ADVANCE website (<http://www.advance-vaccines.eu/>) or using <http://apps.p-95.com/VEMisclassification/>.

It is well-known that exposure- and disease misclassification might jeopardize the validity of VE studies and that such studies require careful design. The simulation tool might help researchers to anticipate at design stage the magnitude and direction of the bias when estimating VE based on potentially misclassified data. As such, this tool can guide the selection of the exposure- and disease definitions that will minimize bias due to misclassification. In addition, if the potential impact of misclassification is found to be unacceptable, several methods to adjust estimates for misclassification exist, although they are not yet commonly used in pharmacoepidemiology [23]. We provided the correction equations for VE estimates in case of non-differential single source (either exposure or disease) misclassification (Section 2.2). Other correction methods include amongst others probabilistic bias analyses [24, 25], Bayesian bias analyses [26–28], modified maximum likelihood methods [29] and imputation-like methods [30–33]. All these methods require assumptions on or estimates of the disease- and exposure misclassification parameters, which—if deemed required—can be obtained using internal or external validation studies.

Several limitations or areas of further development are worth considering. The simulation tool singles out the impact of disease- and exposure misclassification and ignores other sources of bias. Specifically, it is assumed that there is no confounding and no selection bias. In addition, the tool does not include dependent misclassification. For binary variables, misclassification is dependent when the probability of misclassification of one variable depends on the correctness of classification on the other variable [34]. Dependent measurement errors might arise, for example, if data on both exposure and outcome were obtained from medical records with data paucity for some but not all subjects. Furthermore, the tool assumes binary disease-

and exposure variables, whereas particularly the exposure variable might be polytomous (no vaccination, partial or complete vaccination).

The results presented in this paper and the simulation tool may be useful to guide researchers to better design, conduct and interpret future VE studies when data are subject to misclassification. We advocate to use such a simulation tool and modify the parameters according to the study specifics since we have shown that the impact of misclassification strongly depends on the study scenario.

## Author Contributions

**Conceptualization:** Tom De Smedt, Elizabeth Merrall, Denis Macina, Silvia Perez-Vilar, Nick Andrews, Kaatje Bollaerts.

**Formal analysis:** Tom De Smedt, Elizabeth Merrall, Kaatje Bollaerts.

**Methodology:** Tom De Smedt, Elizabeth Merrall, Nick Andrews, Kaatje Bollaerts.

**Software:** Tom De Smedt.

**Supervision:** Kaatje Bollaerts.

**Visualization:** Tom De Smedt.

**Writing – original draft:** Kaatje Bollaerts.

**Writing – review & editing:** Tom De Smedt, Elizabeth Merrall, Denis Macina, Silvia Perez-Vilar, Nick Andrews, Kaatje Bollaerts.

## References

1. Halloran ME, Haber M, Longini IM Jr, Struchiner CJ. Direct and indirect effects in vaccine efficacy and effectiveness. *American journal of epidemiology*. 1991; 133(4):323–31. Epub 1991/02/25. PMID: [1899778](#).
2. Jurek AM, Greenland S, Maldonado G. How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null? *International journal of epidemiology*. 2008; 37(2):382–5. PMID: [18184671](#).
3. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *American journal of epidemiology*. 1977; 105(5):488–95. PMID: [871121](#).
4. Ozasa K. The effect of misclassification on evaluating the effectiveness of influenza vaccines. *Vaccine*. 2008; 26(50):6462–5. <http://dx.doi.org/10.1016/j.vaccine.2008.06.039>. PMID: [18573297](#)
5. Gibbons CL, Mangen MJ, Plass D, Havelaar AH, Brooke RJ, Kramarz P, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*. 2014; 14:147. Epub 2014/02/13. <https://doi.org/10.1186/1471-2458-14-147> PMID: [24517715](#).
6. Flegal KM, Brownie C, Haas JD. The effects of exposure misclassification on estimates of relative risk. *American journal of epidemiology*. 1986; 123(4):736–51. PMID: [3953551](#).
7. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of clinical epidemiology*. 2005; 58(4):323–37. <https://doi.org/10.1016/j.jclinepi.2004.10.012> PMID: [15862718](#).
8. McNeil MM, Gee J, Weintraub ES, Belongia EA, Lee GM, Glanz JM, et al. The Vaccine Safety Datalink: successes and challenges monitoring vaccine safety. *Vaccine*. 2014; 32(42):5390–8. Epub 2014/08/12. <https://doi.org/10.1016/j.vaccine.2014.07.073> PMID: [25108215](#).
9. Nguyen M, Ball R, Midhun K, Lieu TA. The Food and Drug Administration's Post-Licensure Rapid Immunization Safety Monitoring program: strengthening the federal vaccine safety enterprise. *Pharmacoepidemiol Drug Saf*. 2012; 21 Suppl 1:291–7. Epub 2012/01/25. <https://doi.org/10.1002/pds.2323> PMID: [22262619](#).
10. Orenstein EW, De Serres G, Haber MJ, Shay DK, Bridges CB, Gargiullo P, et al. Methodologic issues regarding the use of three observational study designs to assess influenza vaccine effectiveness. *International journal of epidemiology*. 2007; 36(3):623–31. <https://doi.org/10.1093/ije/dym021> PMID: [17403908](#).

11. Jackson ML, Rothman KJ. Effects of imperfect test sensitivity and specificity on observational studies of influenza vaccine effectiveness. *Vaccine*. 2015; 33(11):1313–6. <https://doi.org/10.1016/j.vaccine.2015.01.069> PMID: 25659280.
12. Hanquet G, Valenciano M, Simondon F, Moren A. Vaccine effects and impact of vaccination programmes in post-licensure studies. *Vaccine*. 2013; 31(48):5634–42. Epub 2013/07/17. <https://doi.org/10.1016/j.vaccine.2013.07.006> PMID: 23856332.
13. Foppa IM, Haber M, Ferdinands JM, Shay DK. The case test-negative design for studies of the effectiveness of influenza vaccine. *Vaccine*. 2013; 31(30):3104–9. <https://doi.org/10.1016/j.vaccine.2013.04.026> PMID: 23624093.
14. Farrington CP. Estimation of vaccine effectiveness using the screening method. *Int J Epidemiol*. 1993; 22(4):742–6. Epub 1993/08/01. PMID: 8225751.
15. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, editor. Vienna, Austria 2013.
16. Winston Chang JC, JJ Allaire, Yihui Xie and Jonathan McPherson. Shiny: Web Application Framework for R. 2016.
17. Mereckiene J, Cotter S, Nicoll A, Lopalco P, Noori T, Weber J, et al. Seasonal influenza immunisation in Europe. Overview of recommendations and vaccination coverage for three seasons: pre-pandemic (2008/09), pandemic (2009/10) and post-pandemic (2010/11). *Euro Surveill*. 2014; 19(16):20780. PMID: 24786262.
18. Zhang L, Prietsch SO, Axelsson I, Halperin SA. Acellular vaccines for preventing whooping cough in children. *Cochrane Database Syst Rev*. 2014;(9):CD001478. <https://doi.org/10.1002/14651858.CD001478.pub6> PMID: 25228233.
19. Gustafsson L, Hallander HO, Olin P, Reizenstein E, Storsaeter J. A controlled trial of a two-component acellular, a five-component acellular, and a whole-cell pertussis vaccine. *N Engl J Med*. 1996; 334(6):349–55. <https://doi.org/10.1056/NEJM199602083340602> PMID: 8538705.
20. Stehr K, Cherry JD, Heining U, Schmitt-Grohe S, uberall M, Laussucq S, et al. A comparative efficacy trial in Germany in infants who received either the Lederle/Takeda acellular pertussis component DTP (DTaP) vaccine, the Lederle whole-cell component DTP vaccine, or DT vaccine. *Pediatrics*. 1998; 101(1 Pt 1):1–11. PMID: 9417143.
21. WHO/UNICEF. Estimates of national immunization coverage. DTP3 & DTP4. [07/06/2016]. [http://apps.who.int/immunization\\_monitoring/globalsummary/timeseries/tswucoveragebcg.htm](http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tswucoveragebcg.htm).
22. De Serres G, Skowronski DM, Wu XW, Ambrose CS. The test-negative design: validity, accuracy and precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical trials. *Euro Surveill*. 2013; 18(37). PMID: 24079398.
23. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep*. 2014; 1(4):175–85. <https://doi.org/10.1007/s40471-014-0027-z> PMID: 26085977.
24. Lash TL F M.P.; Fink A.K. Applying Quantitative Bias Analysis to Epidemiologic Data. New York: Springer; 2009.
25. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International journal of epidemiology*. 2005; 34(6):1370–6. Epub 2005/09/21. <https://doi.org/10.1093/ije/dyi184> PMID: 16172102.
26. Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. Hall Ca, editor 2003.
27. MacLehose RF, Olshan AF, Herring AH, Honein MA, Shaw GM, Romitti PA, et al. Bayesian methods for correcting misclassification: an example from birth defects epidemiology. *Epidemiology*. 2009; 20(1):27–35. <https://doi.org/10.1097/EDE.0b013e31818ab3b0> PMID: 19234399.
28. Goldstein ND, Burstyn I, Newbern EC, Tabb LP, Gutowski J, Welles SL. Bayesian Correction of Misclassification of Pertussis in Vaccine Effectiveness Studies: How Much Does Underreporting Matter? *American journal of epidemiology*. 2016; 183(11):1063–70. Epub 2016/05/18. <https://doi.org/10.1093/aje/kwv273> PMID: 27188939.
29. Lyles RH, Tang L, Superak HM, King CC, Celentano DD, Lo Y, et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology*. 2011; 22(4):589–97. PMID: 21487295.
30. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International journal of epidemiology*. 2006; 35(4):1074–81. <https://doi.org/10.1093/ije/dyl097> PMID: 16709616.
31. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *American journal of epidemiology*. 2013; 177(9):904–12. <https://doi.org/10.1093/aje/kws340> PMID: 24627573.

32. Kuchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*. 2006; 62(1):85–96. <https://doi.org/10.1111/j.1541-0420.2005.00396.x> PMID: 16542233.
33. Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in medicine*. 2008; 27(30):6332–50. <https://doi.org/10.1002/sim.3458> PMID: 18937275.
34. Hernan MA, Cole SR. Invited Commentary: Causal diagrams and measurement bias. *American journal of epidemiology*. 2009; 170(8):959–62; discussion 63–4. <https://doi.org/10.1093/aje/kwp293> PMID: 19755635.