
BIAS CORRECTION METHODS FOR TEST-NEGATIVE DESIGNS IN THE PRESENCE OF MISCLASSIFICATION

Akira Endo^{1*}, Sebastian Funk^{1,2}, Adam J. Kucharski^{1,2}

¹ Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine
Keppel St., London WC1E 7HT, The United Kingdom

² Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine
Keppel St., London WC1E 7HT, The United Kingdom

ABSTRACT

The test-negative design has become a standard approach for vaccine effectiveness studies. However, previous studies suggested that it may be more sensitive than other designs to misclassification of disease outcome caused by imperfect diagnostic tests. This could be a particular limitation in vaccine effectiveness studies where simple tests (e.g. rapid influenza diagnostic tests) are used for logistical convenience. To address this issue, we derived a mathematical representation of the test-negative design with imperfect tests, then developed a bias correction framework for possible misclassification. Test-negative design studies usually include multiple covariates other than vaccine history to adjust potential confounders; our methods can also address multivariate analyses and be easily coupled with existing estimation tools. We validated the performance of these methods using simulations of common scenarios for vaccine efficacy and were able to obtain unbiased estimates in a variety of parameter settings.

Keywords Test negative design · Vaccine effectiveness · Misclassification · Bias correction · Statistical methods · Sensitivity · Specificity

Abbreviations VE, vaccine effectiveness; TND, test-negative design; TD, target disease; ND, non-target disease; PCR, polymerase chain reaction; MLE, maximum likelihood estimate; MO, multiple overimputation; EM, expectation-maximisation.

1 Introduction

Vaccine effectiveness (VE) is typically estimated as the vaccine-induced risk reduction of the target disease (TD) and has been traditionally studied by the cohort or case-control designs. However, the test-negative design (TND) is becoming a popular alternative design for vaccine effectiveness (VE) studies [1, 2]. This is a modified version of the case-control study with an alternative definition of the control group; traditional case-control studies usually define controls as non-disease individuals in the study population, while the TND studies use individuals with similar symptoms to the target disease but presenting negative test results (i.e., patients of non-target diseases; ND). The test-negative design can therefore minimise ascertainment bias by including only medically-attended patients in both case and control groups. Many TND studies have focused on influenza vaccination, but recent studies have also targeted other diseases including pneumococcal disease [3, 4] and rotavirus disease [5, 6, 7].

Despite its increasing popularity, TND can be more sensitive than other study designs to misclassification of disease outcome. Multiple studies have shown that VE is underestimated when the diagnostic tests used in the study are imperfect (i.e. have a sensitivity and/or a specificity less than 100%) [8, 9, 10]. This can be a particular issue when simple tests (e.g. rapid diagnostic tests) are used for logistical convenience, as simple tests tend to have lower diagnostic

*Correspondence to: akira.endo@lshtm.ac.uk

15 performance than more advanced tests (e.g. polymerase chain reaction; PCR). Previous studies evaluated the expected
 16 degree of bias and concluded that specificity had a more important effect on bias than sensitivity [8, 9, 10]. These
 17 findings appear to support the use of rapid tests, despite limited sensitivity, because the specificity of these tests
 18 is typically high [2]. However, theoretical studies to date have been based on specific assumptions about efficacy
 19 and pathogen epidemiology; it is therefore unclear whether such conclusions hold for all plausible combinations of
 20 scenarios.

21 If a study is expected to generate a non-negligible bias in estimation, such bias needs to be assessed and—if possible—
 22 corrected before the estimate is reported. Greenland [11] proposed a bias correction method for cohort studies where the
 23 sensitivity and specificity of the test are known (or at least assumed). However, this method cannot apply to case-control
 24 studies because of differential recruitment, whereby the probability of recruiting (test-positive) cases and (negative)
 25 controls may be different. Although TND studies are often considered to be special cases of case-control studies, they
 26 are free from the issue of differential recruitment because the recruitment and classification are mutually-independent.
 27 This means that, while Greenland’s method does not apply to TND, another type of bias correction may still be possible.
 28 For example, De Smedt et al. have characterised the misclassification bias in VE in the test-negative design in a
 29 simulation study [3]. One limitation of this approach is it relies on the unobserved “true” disease risk being known,
 30 where in reality this is not usually measurable in field studies. As a result, bias correction methods for TND studies that
 31 are directly applicable to field data have not yet been proposed. Moreover, previous analysis of misclassification bias
 32 has not considered the impact of multivariate analysis, where potential confounders (e.g. age and sex) are also included
 33 in the model used to estimate VE.

34 To address these issues, we develop a bias correction method for the test-negative VE studies that uses only data
 35 commonly available in field studies. We also apply these methods to multivariate analyses. As our approach uses the
 36 so-called multiple overimputation framework (generalisation of multiple imputation) [12], it can easily be coupled with
 37 a wide range of estimation tools without modifying their inside algorithms. Finally, we evaluate the performance of our
 38 methods by simulations of plausible epidemiological scenarios.

39 2 Methods and results

40 2.1 Characterising bias in test negative design studies

41 First, we consider the case where only vaccination history is included as a risk factor of acquiring the TD (i.e. the
 42 univariate setting). In this case, the (true) case counts can in theory be summarised in a two-by-two table as shown
 43 below:

	Vaccinated	Unvaccinated
Target disease	x_V	x_U
Non-target disease	y_V	y_U
Subtotal	S_V	S_U

44 Following the approach of Haber et al. (2015) [13], we consider four steps in the case reporting process: vaccination,
 45 onset of symptoms, seeking of medical care, and diagnosis. For simplicity, let us assume that occurrence of TD and ND
 46 are mutually independent, where their prevalences are represented as r_1 and r_0 , respectively¹. Let v be the vaccination
 47 coverage; in observational studies, vaccinated and unvaccinated population can have different likelihoods of seeking
 48 medical treatment (we denote such probabilities as m_V and m_U , respectively). As our focus in the present study is the
 49 bias in VE estimation caused by imperfect tests, we made two key assumptions following Haber et al [13]: vaccination
 50 does not affect the risk of ND or the relative probability μ of medical attendance between TD and ND (which may
 51 reflect different disease severity between TD and ND). Namely, the study was assumed to be able to provide an unbiased
 52 VE estimate if tests are perfect. Based on these assumptions, we can classify the expected incidence in population N
 53 into four categories:

	Vaccinated	Unvaccinated
Target disease	$v\gamma\mu m_V r_1 N$	$(1-v)\mu m_U r_1 N$
Non-target disease	$v m_V r_0 N$	$(1-v)m_U r_0 N$
Subtotal	$v m_V (\gamma\mu r_1 + r_0) N$	$(1-v)m_U (\mu r_1 + r_0) N$

¹It has been suggested that a possible violation of this assumption occur as a result of virus interference [14], but conclusive evidence for this is currently lacking [15, 16] and the effect on VE estimates may be limited in any case [17]

54 where γ is the relative risk of TD in the vaccinated population (i.e., $\gamma = 1 - \text{VE}$). m_V and m_U are the probabilities of
 55 the vaccinated and unvaccinated seeking medical care given ND (those given TD are μm_V and μm_U , respectively). In
 56 TND studies, the (true) odds ratio corresponds to the relative risk γ .

57 Suppose that the true data in a TND study (x_V, y_V, x_U, y_U) is as described in the above tables. However, due to
 58 imperfect tests, we would instead expect to obtain the following observations:

	Vaccinated	Unvaccinated
Test positive	$\alpha x_V + (1 - \beta)y_V$	$\alpha x_U + (1 - \beta)y_U$
Test negative	$(1 - \alpha)x_V + \beta y_V$	$(1 - \alpha)x_U + \beta y_U$
Subtotal	S_V	S_U

59 where α and β are the sensitivity and specificity of the test, respectively. Denoting observed case counts with
 60 misclassification by X and Y , the process of diagnosis can be represented by the following matrix expression:

$$\begin{bmatrix} X_V & X_U \\ Y_V & Y_U \end{bmatrix} = \begin{bmatrix} \alpha & 1 - \beta \\ 1 - \alpha & \beta \end{bmatrix} \begin{bmatrix} x_V & x_U \\ y_V & y_U \end{bmatrix}. \quad (1)$$

61 Matrix $C = \begin{bmatrix} \alpha & 1 - \beta \\ 1 - \alpha & \beta \end{bmatrix}$ describes the conversion from the true disease state to the observed result. We hereafter
 62 refer to C as the classification matrix.

63 The observed odds ratio (subject to the misclassification bias) is therefore given as

$$\begin{aligned} \frac{X_V}{Y_V} / \frac{X_U}{Y_U} &= \frac{\alpha x_V + (1 - \beta)y_V}{(1 - \alpha)x_V + \beta y_V} / \frac{\alpha x_U + (1 - \beta)y_U}{(1 - \alpha)x_U + \beta y_U} \\ &= \frac{[\alpha \gamma \mu r_1 + (1 - \beta)r_0][\alpha \mu r_1 + \beta r_0]}{[(1 - \alpha)\gamma \mu r_1 + \beta r_0][\alpha \mu r_1 + (1 - \beta)r_0]} \\ &= \frac{[\alpha \gamma \delta + (1 - \beta)][(1 - \alpha)\delta + \beta]}{[(1 - \alpha)\gamma \delta + \beta][\alpha \delta + (1 - \beta)]}, \end{aligned} \quad (2)$$

64 where $\delta = \frac{r_1 \mu}{r_0}$ is the odds of the (medically-attended) target disease in the unvaccinated population.

65 We define bias in the VE estimate to be the absolute difference between the (raw) estimate and the true value. The
 66 expected bias B is a function of α, β, γ and δ :

$$\begin{aligned} B(\alpha, \beta, \gamma, \delta) &= \text{VE}_{\text{raw}} - \text{VE}_{\text{true}} \\ &= \gamma - \frac{[\alpha \gamma \delta + (1 - \beta)][(1 - \alpha)\delta + \beta]}{[(1 - \alpha)\gamma \delta + \beta][\alpha \delta + (1 - \beta)]}. \end{aligned} \quad (3)$$

67 This suggests that the influence of sensitivity/specificity on the degree of bias varies depending on the case ratio
 68 $\delta/(1 + \delta)$, i.e. the ratio between incidence of medical attendance for TD and ND in the study population (Figure 1).
 69 The degree of bias also depends on γ but is independent of m_V and m_U . The degree of bias is largely determined by
 70 the test specificity when the case ratio is small, but the influence of sensitivity and specificity is almost equivalent with a
 71 case ratio of 0.6. It is notable that high specificity does not always assure that the bias is negligible. This may be true if
 72 specificity is strictly 100% and the case ratio is low to moderate, but a slight decline to 97% can cause a bias up to
 73 10-15 percentage points. The effect of sensitivity is also non-negligible when the case ratio is high.

74 When the expected bias is plotted against the case ratio with various combinations of test performance, we find that
 75 VE estimates can be substantially biased for certain case ratios (especially when the ratio is far from 1:1), even with
 76 reasonably high sensitivity and specificity (Figure 2A). In TND studies, researchers have no control over the case ratio
 77 because the study design requires that all tested individuals be included in the study. We found that the proportion of
 78 TD-positive patients in previous TND studies (retrieved from three systematic reviews [18, 19, 20]) varied considerably,
 79 ranging from 10% to 70% (Figure 2B)². Because of this large variation in the case ratio, it would be difficult to
 80 predict the degree of bias before data collection. Post-hoc assessment and correction therefore need to be considered.
 81 (Further analysis of the relationship between the degree of bias and parameter values can be found in the supplementary
 82 information.)

²Strictly speaking, proportion positive is a different quantity from case ratio, but it should serve as a reasonable proxy of the case ratio in most settings

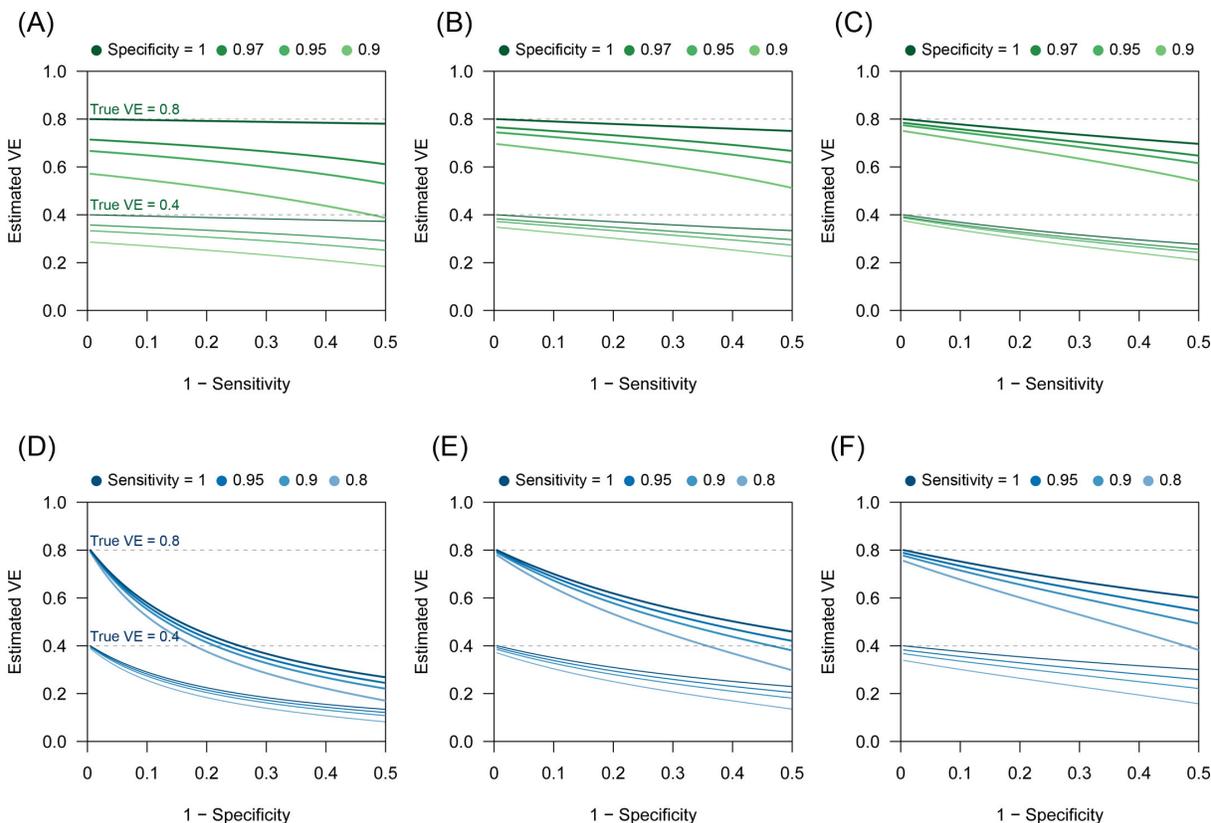


Figure 1: Bias in VE estimates caused by misclassification for different combinations of parameter values. (A) – (C) Estimated VE plotted against sensitivity. (A) True case ratio = 0.2 (B) 0.4 (C) 0.6. Each two sets of lines respectively correspond to different true VEs (80% and 40%, denoted by the dotted lines). (D) – (F) Estimated VE plotted against specificity. (D) True case ratio = 0.2 (E) 0.4 (F) 0.6.

83 2.2 Bias correction in univariate analysis

84 2.2.1 Model and statistical analysis

85 To develop a correction method that can address the bias presented in the previous section, we first model the case
 86 reporting process in the univariate setting as follows. Let us assume that incidence of TD and ND both follow
 87 Poisson-distributions. As presented in Section 2.1, the mean total incidence in the unvaccinated population is given
 88 as $\lambda_U = (1 - v)m_U(r_1\mu + r_0)N$. Let $\lambda_V = \frac{vm_V}{(1-v)m_U}\lambda_U$ so that λ_V corresponds to the mean total incidence in
 89 the vaccinated population ($= Nv$) when $\gamma = 1$, i.e. $VE=0$. This definition is to ensure that parameters γ and λ_V
 90 are mutually independent. Let $\delta = \frac{r_1\mu}{r_0}$ be the odds of the (medically-attended) target disease in the unvaccinated
 91 population. Using these four parameters $\gamma, \delta, \lambda_V, \lambda_U$, we get the following table for (potentially mis-classified) mean
 92 case counts:

	Vaccinated	Unvaccinated
Test positive	$\frac{\alpha\gamma\delta+(1-\beta)}{1+\delta}\lambda_V$	$\frac{\alpha\delta+(1-\beta)}{1+\delta}\lambda_U$
Test negative	$\frac{(1-\alpha)\gamma\delta+\beta}{1+\delta}\lambda_V$	$\frac{(1-\alpha)\delta+\beta}{1+\delta}\lambda_U$
Subtotal	$\frac{1+\gamma\delta}{1+\delta}\lambda_V$	λ_U

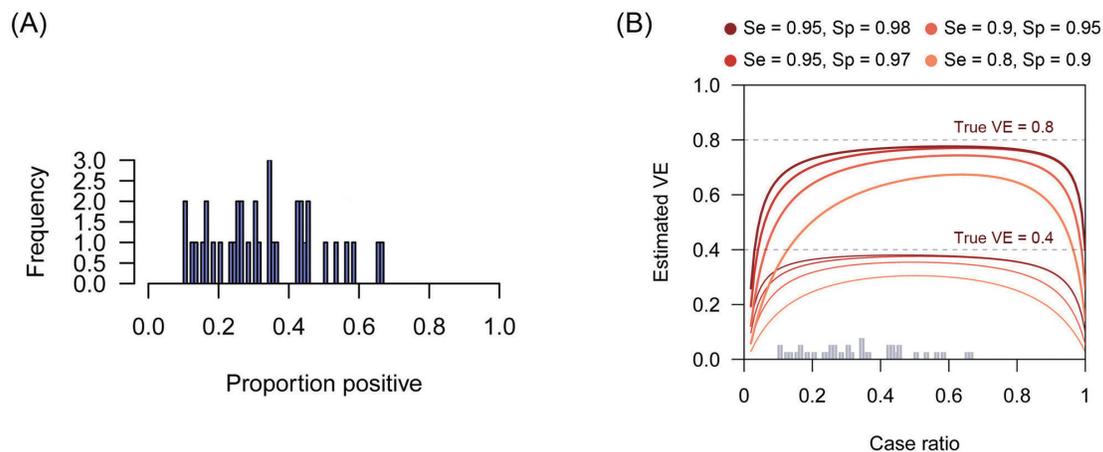


Figure 2: Biased VE estimates with varying case ratio and the observed proportion of positive patients. (A) The proportion of test positive patients in TND studies from systematic reviews. The proportions were retrieved from three systematic reviews [18, 19, 20]. (B) Estimated VE plotted against case ratio. Two sets of lines respectively correspond to different true VEs (80% and 40%, denoted by the dotted lines). The histogram in Panel (A) is overlaid on the x axis.

93 When data $D = (X_V, Y_V, X_U, Y_U)$ is obtained following this misclassified pattern, we can construct the likelihood of
 94 obtaining such data, given underlying parameters, as

$$\begin{aligned} \mathcal{L}(\gamma, \delta, \lambda_V, \lambda_U; D) = & \text{Pois} \left(X_V; \frac{\alpha\gamma\delta + (1-\beta)}{1+\delta} \lambda_V \right) \text{Pois} \left(Y_V; \frac{(1-\alpha)\gamma\delta + \beta}{1+\delta} \lambda_V \right) \\ & \text{Pois} \left(X_U; \frac{\alpha\delta + (1-\beta)}{1+\delta} \lambda_U \right) \text{Pois} \left(Y_U; \frac{(1-\alpha)\delta + \beta}{1+\delta} \lambda_U \right). \end{aligned} \quad (4)$$

95 By maximising this likelihood over all parameters, we can obtain a maximum likelihood estimate (MLE) of the odds
 96 ratio γ^* that accounts for misclassification. Let us refer to

$$\gamma^* = \frac{X_V - \frac{1-\beta}{\beta} Y_V}{Y_V - \frac{1-\alpha}{\alpha} X_V} \cdot \frac{Y_U - \frac{1-\alpha}{\alpha} X_U}{X_U - \frac{1-\beta}{\beta} Y_U} \quad (5)$$

97 as the "corrected odds ratio", which gives an unbiased estimate of γ . Comparing γ^* with the the "raw" odds ratio $\frac{X_V Y_U}{Y_V X_U}$,
 98 we find that the estimate can be corrected using the following substitution

$$\begin{aligned} X_V & \rightarrow X_V - o_\beta Y_V \\ Y_V & \rightarrow Y_V - o_\alpha X_V \\ X_U & \rightarrow X_U - o_\beta Y_U \\ Y_U & \rightarrow Y_U - o_\alpha X_U \end{aligned} \quad (6)$$

99 where $o_\alpha = \frac{1-\alpha}{\alpha}$ and $o_\beta = \frac{1-\beta}{\beta}$ are the odds of diagnostic errors corresponding to sensitivity and specificity,
 100 respectively, which take 0 when sensitivity/specificity is perfect. Also note that the same odds ratio is obtained by
 101 taking the odds ratio of the reconstructed data table where the inverted classification matrix is applied:

$$C^{-1} \begin{bmatrix} X_V & X_U \\ Y_V & Y_U \end{bmatrix} = \frac{1}{\alpha + \beta - 1} \begin{bmatrix} \beta & 1 - \beta \\ 1 - \alpha & \alpha \end{bmatrix} \begin{bmatrix} X_V & X_U \\ Y_V & Y_U \end{bmatrix}. \quad (7)$$

102 The determinant $c = \alpha + \beta - 1$ is the Youden index of the test and satisfies $0 < c \leq 1$ (if $c < 0$, the test is not predictive
 103 and the definitions of positive/negative should be swapped).

104 In some (relatively rare) cases, one or more quantities in Eq. (6) may become negative due to random fluctuations in
 105 observation. Theoretically, negative values are not permitted in as true case counts, and thus such negative quantities
 106 would need to be truncated to 0. As a result, the corrected odds ratio can be either 0 or infinity. Such an estimate would
 107 suggest that the study does not have a sufficient sample size to properly evaluate VE and that the study design itself
 108 might need to be reconsidered.

The confidence interval for VE can be obtained by assuming log-normality of the odds ratio γ , i.e.,

$$\gamma = \gamma^* \exp(\pm 1.96\sigma^*),$$

109 where σ is the shape parameter of the log-normal distribution and is empirically given as

$$\begin{aligned} \sigma^* &= \text{SD}(\log(\gamma^*)) \\ &= c \sqrt{\frac{1}{S_V} \cdot \frac{\pi_V(1-\pi_V)}{(1-\pi_V-(1-\alpha))^2(\pi_V-(1-\beta))^2} + \frac{1}{S_U} \cdot \frac{\pi_U(1-\pi_U)}{(1-\pi_U-(1-\alpha))^2(\pi_U-(1-\beta))^2}}, \end{aligned} \quad (8)$$

110 where π_V and π_U are observed (uncorrected) TD frequency ($\pi_V = \frac{X_V}{X_V+Y_V}$ and $\pi_U = \frac{X_U}{X_U+Y_U}$). (See Appendix for
 111 details of the MLE and confidence intervals.)

112 2.2.2 Simulation

113 To assess the performance of the corrected odds ratio given in Equation (22), we used simulation studies. TND
 114 study datasets were generated based on the likelihood presented in Equation (4), where the mean total sam-
 115 ple size ($\frac{1+\gamma\delta}{1+\delta}\lambda_V + \lambda_U$) was set to be 3,000. Parameter values were chosen according to a range of scenar-
 116 ios shown in Table 1, and the true vaccine effectiveness $\text{VE} = 1 - \gamma$ was compared with the estimates ob-
 117 tained from the simulated data. For each scenario, simulation was repeated 500 times to yield the distribution
 118 of estimates. Reproducible codes (including those for simulations in later sections) are reposted on GitHub
 119 (<https://github.com/akira-endo/TND-biascorrection/>).

Table 1: Simulation settings

ID	Scenario	True VE (γ)	λ_V/λ_U	Case ratio ($\frac{\gamma}{1+\gamma}$)	Sensitivity (α)	Specificity (β)
1	Baseline: low VE	0.4	0.5	0.5	0.8	0.95
2	Baseline: high VE	0.8	0.5	0.5	0.8	0.95
3	High quality test: low VE	0.4	0.5	0.5	0.95	0.97
4	High quality test: high VE	0.8	0.5	0.5	0.95	0.97
5	Low quality test: low VE	0.4	0.5	0.5	0.6	0.9
6	Low quality test: high VE	0.8	0.5	0.5	0.6	0.9
7	High TD incidence: low VE	0.4	0.5	0.7	0.8	0.95
8	High TD incidence: high VE	0.8	0.5	0.7	0.8	0.95
9	Low TD incidence: low VE	0.4	0.5	0.3	0.8	0.95
10	Low TD incidence: high VE	0.8	0.5	0.3	0.8	0.95
11	High vaccine coverage: low VE	0.4	0.7	0.5	0.8	0.95
12	High vaccine coverage: high VE	0.8	0.7	0.5	0.8	0.95
13	Low vaccine coverage: low VE	0.4	0.3	0.5	0.8	0.95
14	Low vaccine coverage: high VE	0.8	0.3	0.5	0.8	0.95

120 We found that the uncorrected estimates, directly obtained from the raw case counts that were potentially misclassified,
 121 exhibited substantial underestimation of VE for most parameter values (Figure 3). On the other hand, our bias correction
 122 method was able to yield unbiased estimates in every setting, whose median almost correspond to the true VE. Although
 123 the corrected and uncorrected distributions were similar (with a difference in median $\sim 5\%$) when VE is relatively
 124 low (40%) and the test has sufficiently high sensitivity and specificity (95% and 97%, respectively), they became
 125 distinguishable with a higher VE (80%). With lower test performances, the bias in the VE estimates can be up to
 126 10-20%, which may be beyond the level of acceptance in VE studies.

127 2.2.3 Bias correction of VEs reported in previous studies

128 We have seen that the degree of bias for uncorrected VE estimates depends on parameter values. To explore the possible
 129 degree of bias in existing VE studies, we extracted the reported crude VEs (i.e. VEs without adjustments of potential
 130 confounders) from two systematic reviews [18, 20] (Young et al. [19] was not included because they did not report
 131 case counts) and applied our bias correction method assuming different levels of test sensitivity and specificity. The

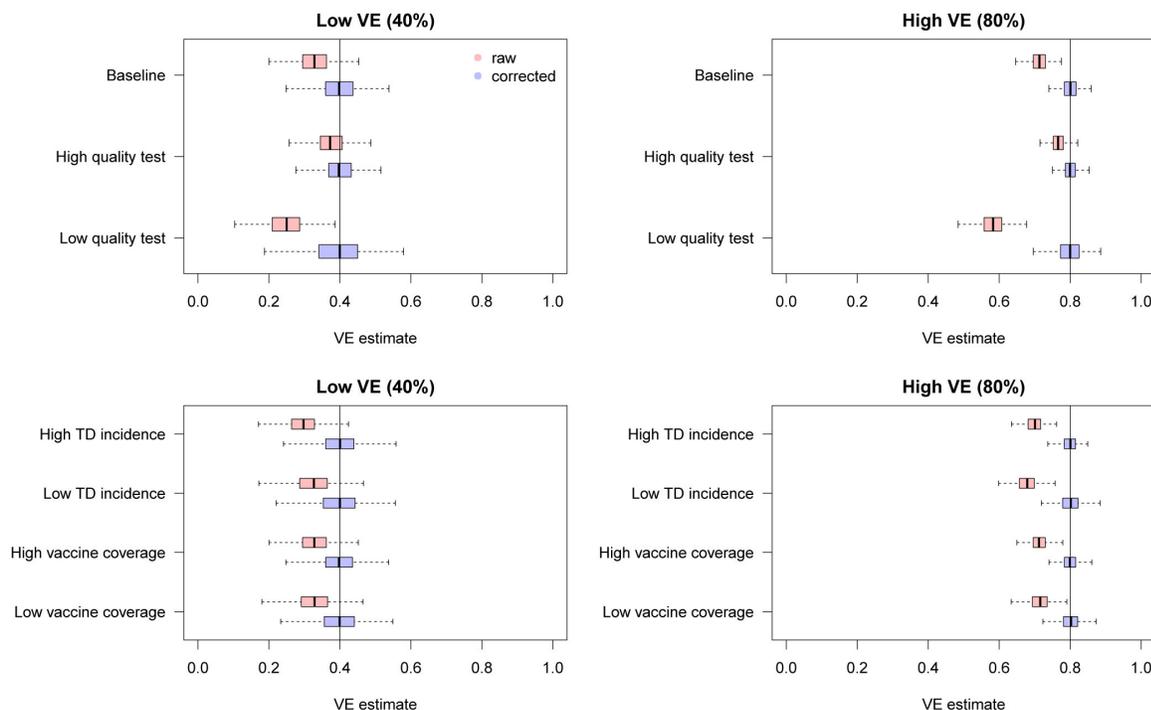


Figure 3: Bias correction for simulated data in the univariate setting. The distributions of bias-corrected VE estimates (boxplots in blue) are compared with those of raw VE estimates without correction (red). Five hundred independent datasets were randomly generated for each set of parameter values, and the corrected and uncorrected VE estimates are compared with the true value (black solid line).

132 case counts for each study summarised in the reviews were considered eligible for the analysis if the total sample size
 133 exceeded 200. Varying the assumed sensitivity and specificity, we investigated the possible discrepancy between the
 134 reported VE (or crude VE derived from the case counts if unreported in the reviews) and bias-corrected VE. We did not
 135 consider correcting adjusted VEs because it requires access to the original datasets.

136 Figure 4 displays the discrepancy between the reported VE and bias corrected VE corresponding to a range of
 137 assumptions on the test performance. Many of the extracted studies employed polymerase chain reaction (PCR)
 138 for the diagnostic test, which is expected to have a high performance. However, the true performance of PCR
 139 cannot be definitively measured as there is currently no other gold-standard test available. Figure 4B suggests that
 140 even a slight decline in the test performance can introduce a non-negligible bias in some parameter settings. Our
 141 bias correction methods may therefore also be useful in TND studies using PCR, which would enable a sensitivity
 142 analysis accounting for potential misdiagnosis by PCR tests. In this light, it is useful that the corrected odds ratio
 143 $\gamma^* = \frac{X_V - \frac{1-\beta}{\beta} Y_V}{Y_V - \frac{1-\alpha}{\alpha} X_V} \cdot \frac{Y_U - \frac{1-\alpha}{\alpha} X_U}{X_U - \frac{1-\beta}{\beta} Y_U}$ is a monotonic function of both α and β (given that all the four components are positive).
 144 The possible range of VE in a sensitivity analysis is obtained by supplying γ^* with the assumed upper and lower limits
 145 of sensitivity and specificity.

146 2.3 Bias correction in multivariate analysis

147 2.3.1 Theoretical framework

148 TND studies often employ a multivariate regression framework to address potential confounding variables such as
 149 age. The most widespread approach is to use linear models (e.g., logistic regression) and include vaccination history
 150 as well as other confounding variables as covariates. The estimated linear coefficient for vaccination history can

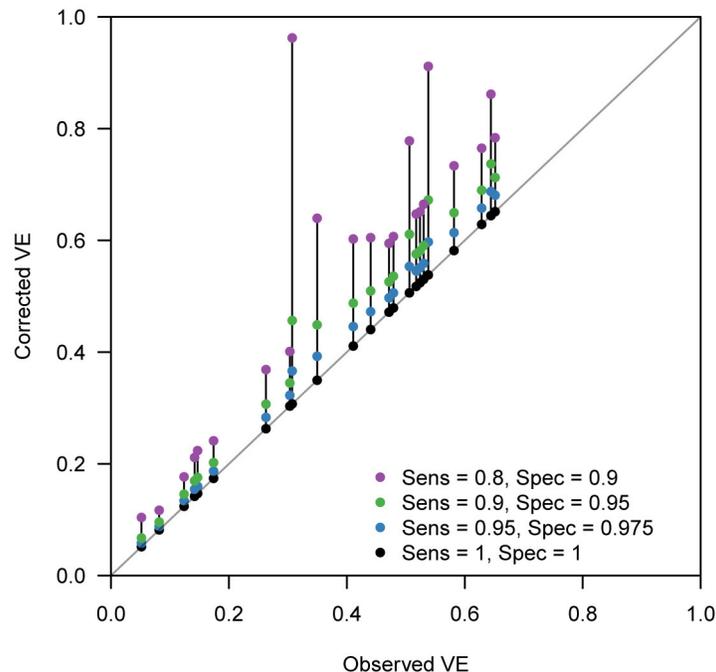


Figure 4: Bias correction method applied to published VE estimates assuming various test sensitivity and specificity. Case count data were extracted from two systematic reviews [18, 20]. Each connected set of dots show how (crude) VE estimates reported in the review varies when imperfect sensitivity and specificity are assumed. Black dots on the grey diagonal line denote the original VEs reported in the reviews (where sensitivity = specificity = 1) and coloured dots show the estimated VE considering potential misclassification.

151 then be converted VE (in the logistic regression model, the linear coefficient for vaccination history corresponds to
 152 $\log(1 - VE)$).

153 In this situation, the likelihood function now reflects a regression model and thus the bias-corrected estimate in the
 154 univariate analysis (Equation (5)) is no longer applicable. We therefore need to develop a separate multivariate TND
 155 study framework to correct for bias in multivariate analysis. Suppose that covariates $\xi = (\xi^1, \xi^2, \dots, \xi^n)$ are included
 156 in the model, and that ξ^1 corresponds to vaccination history (1: vaccinated, 0: unvaccinated). ξ is expected to have a
 157 certain distribution over the total population N , and let us denote the frequency density of covariates ξ by $N(\xi)$, where
 158 $\int N(\xi)d\xi = N$. Let $\rho_1(\xi)$ and $\rho_0(\xi)$ be the conditional probabilities that an individual is included in the study with
 159 TD and ND, respectively, given covariates ξ . Incorporating misclassification, the probability of an individual i with
 160 covariates ξ_i being included and tested positive/negative will be

$$\begin{aligned}\rho_+(\xi_i) &= \alpha\rho_1(\xi_i) + (1 - \beta)\rho_0(\xi_i) \\ \rho_-(\xi_i) &= (1 - \alpha)\rho_1(\xi_i) + \beta\rho_0(\xi_i)\end{aligned}\tag{9}$$

161 Assuming that disease incidences follow Poisson distributions, as in the univariate case, we can obtain the probability
 162 density of observing data $D = \{Z_i, \xi_i\}_{i=1,2,\dots,S}$ (Z_i denotes the test result of individual i) as

$$\mathcal{P}(D) = \text{Pois}(S_+; \lambda_+) \text{Pois}(S_-; \lambda_-) \prod_{i \in \{+\}} \frac{\rho_+(\xi_i)N(\xi_i)}{\lambda_+} \prod_{i \in \{-\}} \frac{\rho_-(\xi_i)N(\xi_i)}{\lambda_-}.\tag{10}$$

163 where λ_+ and λ_- are the mean incidence of being included in the study and tested positive/negative: $\lambda_{\pm} =$
 164 $\int \rho_{\pm}(\xi)N(\xi)d\xi$. The first two Poisson distributions on the right-hand side of Eq. (10) give the probability that
 165 the study yields S_+ positive and S_- negative subjects. The products that follow represent the probability density for
 166 covariates ξ_i observed in the positive/negative group.

167 Suppose that we model this system using a parameter set θ . We could directly model $\rho_1(\xi_i; \theta)$ and $\rho_0(\xi_i; \theta)$; however,
 168 it is often more convenient to model the binomial probability for the true outcome $p_1(\xi_i) = \frac{\rho_1(\xi_i)}{\rho(\xi_i)}$ and $p_0(\xi_i) = \frac{\rho_0(\xi_i)}{\rho(\xi_i)}$,
 169 where $\rho(\xi_i) = \rho_1(\xi_i) + \rho_0(\xi_i) = \rho_+(\xi_i) + \rho_-(\xi_i)$ is the probability density of being included in the study given
 170 covariates ξ_i , because the absolute scale of incidence is rarely of a primary concern. The binomial probabilities for the
 171 respective observed outcomes (with errors) are then given by:

$$\begin{aligned}\pi_+(\xi_i; \theta) &= \alpha p_1(\xi_i; \theta) + (1 - \beta) p_0(\xi_i; \theta) \\ \pi_-(\xi_i; \theta) &= (1 - \alpha) p_1(\xi_i; \theta) + \beta p_0(\xi_i; \theta)\end{aligned}\quad (11)$$

172 Let us use parameter set θ to model the binomial probabilities π_+ (and π_-) and assume that another set of parameters η
 173 (nuisance parameters) characterise $\rho(\xi_i)$. Then our objective is reduced to the estimation of θ and η .

174 Rearranging Equation (10), we get the joint likelihood for θ and η :

$$\mathcal{L}(\theta, \eta; D) = \binom{S}{S_+} \text{Pois}(S; \lambda(\eta)) \prod_{i=1}^S \frac{\rho(\xi_i; \eta) N(\xi_i)}{\lambda(\eta)} \prod_{i=1}^S \pi_{Z_i}(\xi_i; \theta), \quad (12)$$

175 where $\lambda(\eta)$ is the overall mean incidence: $\lambda(\eta) = \int \rho(\xi; \eta) N(\xi) d\xi$. The factor outside the products on the right-hand
 176 side of Eq. (12) is the probability that the study yields S subjects of which S_+ are positives and S_- are negatives. The
 177 first product is the probability density for covariates ξ_i observed in data D , and the second product is the binomial
 178 probabilities for the test results Z_i . When only θ is of our concern, we can obtain the MLE for θ by maximising

$$\mathcal{L}(\theta; D) = \prod_{i=1}^S \pi_{Z_i}(\xi_i; \theta) = \prod_{i \in \{+\}} [\alpha p_1(\xi_i; \theta) + (1 - \beta) p_0(\xi_i; \theta)] \prod_{i \in \{-\}} [(1 - \alpha) p_1(\xi_i; \theta) + \beta p_0(\xi_i; \theta)], \quad (13)$$

179 as θ and η are separate in the likelihood (12). With the estimate θ^* , the VE estimate for an individual with covariates
 180 $\xi^{2:n} = (\xi^2, \xi^3, \dots, \xi^n)$ is given as (1 - odds ratio):

$$\text{VE}(\xi^{2:n}) = 1 - \frac{p_1(\xi^1 = 1, \xi^{2:n}; \theta^*)}{p_0(\xi^1 = 1, \xi^{2:n}; \theta^*)} \bigg/ \frac{p_1(\xi^1 = 0, \xi^{2:n}; \theta^*)}{p_0(\xi^1 = 0, \xi^{2:n}; \theta^*)}. \quad (14)$$

181 2.3.2 Direct likelihood method for the logistic regression model

182 The logistic regression model is well-suited for modelling binomial probabilities p_1 and p_0 . The log-odds ($\log(\frac{p_1}{p_0})$) is
 183 characterised by a linear predictor as:

$$\log\left(\frac{p_1(\xi; \theta)}{p_0(\xi; \theta)}\right) = \theta_0 + \theta_1 \xi^1 + \dots + \theta_n \xi^n. \quad (15)$$

184 In the logistic regression model where covariate ξ_1 indicates vaccination history, the corresponding coefficient θ_1
 185 gives the VE estimate: $\text{VE} = 1 - \exp(\theta_1)$. Due to the assumed linearity, the estimated VE value is common across
 186 individuals regardless of covariates $\xi^{2:n}$.

187 We can employ the direct likelihood method by combining Equations (13) and (15). The usual logistic regression
 188 optimises θ by assuming that the test results follow Bernoulli distributions $Z_i \sim \text{Bernoulli}(p_1(\xi_i; \theta))$ ($Z_i = 1$ for
 189 positive test results and 0 for negative). To correct the misclassification bias, we instead need to use the modified
 190 probabilities given by Eq. (11) to construct the likelihood accounting for diagnostic error, i.e.,

$$Z_i \sim \text{Bernoulli}(\pi_+(\xi_i; \theta)) = \text{Bernoulli}(\alpha p_1(\xi_i; \theta) + (1 - \beta) p_0(\xi_i; \theta)). \quad (16)$$

191 Parameter θ is estimated by directly maximising the probability of observing $\{Z_i\}$ based on Eq. (16)

192 Note that as long as the binomial probability is the modelling target, other type of models (e.g. probit model) could also
 193 be employed under a similar framework.

194 2.3.3 Multiple overimputation method to be combined with existing tools

195 The direct likelihood method presented in the previous section is the most rigorous MLE approach and would therefore
 196 be preferable whenever possible. However, it is often technically-demanding to implement such approaches as it involves
 197 re-defining the likelihood; if we wanted to use existing tools for logistic regression (or other models), for example,
 198 we would need to modify the internal algorithm specifying the likelihood computation. To ensure that researchers
 199 are able to employ correction methods without reprogramming the underlying software algorithms, we also propose
 200 another method, which employs a multiple overimputation (MO) framework [12] to account for misclassification.
 201 Whereas multiple imputation only considers missing values, MO is proposed as a more general concept which includes
 202 overwriting mismeasured values in the dataset by imputation. In our multivariate bias correction method, test results in
 203 the dataset (which are potentially misclassified) are randomly overimputed.

204 Let M be an existing estimation software tool whose likelihood specification cannot be reprogrammed. Given data
 205 $d = \{z_i, \xi_i\}_{i=1,2,\dots,S}$, where z_i denotes the true disease state ($z = 1$ for TD and $z = 0$ for ND), M would be expected
 206 to return at least the following two elements: the point estimate of VE (ε_d) and the predicted binomial probability $\hat{p}_1(\xi_i)$
 207 for each individual i . From the original observed dataset D , imputed datasets $\{\tilde{D}^j\} = \{\tilde{D}^1, \tilde{D}^2, \dots, \tilde{D}^J\}$ are generated
 208 by the following procedure.

- 209 1. For $i = 1, 2, \dots, S$, impute disease state \tilde{z}_i^j based on the test result Z_i . Each \tilde{z}_i^j is sampled from a Bernoulli
 210 distribution conditional to Z_i :

$$\tilde{z}_i^j \sim \begin{cases} \text{Bernoulli}(1 - \tilde{\varphi}_{i+}) & (Z_i = 1) \\ \text{Bernoulli}(\tilde{\varphi}_{i-}) & (Z_i = 0) \end{cases}. \quad (17)$$

211 $\tilde{\varphi}_{i+}$ and $\tilde{\varphi}_{i-}$ are estimated probabilities that the test result for individual i is incorrect (i.e., $z_i \neq Z_i$) given Z_i .
 212 The sampling procedure (17) is therefore interpreted as the test result Z_i being "flipped" at a probability $\tilde{\varphi}_{i+}$
 213 or $\tilde{\varphi}_{i-}$. Later we will discuss possible procedures to obtain these probabilities.

- 214 2. Apply M to $\tilde{D}^j = \{\tilde{z}_i^j, \xi_i\}$ to yield a point estimate of VE (ε^j).
- 215 3. Repeat 1., 2. for $j = 1, 2, \dots, J$ to yield MO estimates $\{\varepsilon^j\}_{j=1,\dots,J}$.

216 Once MO estimates $\{\varepsilon^j\}$ are obtained, the pooled estimate and confidence intervals of VE are obtained by appropriate
 217 summary statistics, e.g., Rubin's rules. As long as the estimated "flipping" probabilities $\tilde{\varphi}_{i\pm} = (\tilde{\varphi}_{i+}, \tilde{\varphi}_{i-})$ are well
 218 chosen, this MO procedure should provide an unbiased estimate of VE with a sufficiently large number of iterations J .

219 There can be multiple candidates for the flipping probability estimate $\tilde{\varphi}_{i\pm}$. Here we discuss two possible options:
 220 parametric bootstrapping and the expectation-maximisation (EM) algorithm. Our simulation showed that parametric
 221 bootstrapping is preferable (see the supplementary document).

222 (i) Parametric bootstrapping

223 The simplest option to estimate $\tilde{\varphi}_{i\pm}$ is to use Bayesian probability

$$\mathcal{P}(z_i = Z_i | Z_i) = \begin{cases} \frac{\alpha \mathcal{P}(z_i = 1)}{\alpha \mathcal{P}(z_i = 1) + (1 - \beta) \mathcal{P}(z_i = 0)} = \frac{\alpha p_1(\xi_i)}{\alpha p_1(\xi_i) + (1 - \beta) p_0(\xi_i)} & (Z_i = 1) \\ \frac{\beta \mathcal{P}(z_i = 0)}{(1 - \alpha) \mathcal{P}(z_i = 1) + \beta \mathcal{P}(z_i = 0)} = \frac{\beta p_0(\xi_i)}{(1 - \alpha) p_1(\xi_i) + \beta p_0(\xi_i)} & (Z_i = 0) \end{cases} \quad (18)$$

224 The true binomial probabilities $p_0(\xi_i)$, $p_1(\xi_i)$ are not known, but their estimators are derived with the inverted
 225 classification matrix in the same manner as Eq. (7). By substituting $\begin{bmatrix} p_1(\xi_i) \\ p_0(\xi_i) \end{bmatrix}$ with $C^{-1} \begin{bmatrix} \pi_+(\xi_i) \\ \pi_-(\xi_i) \end{bmatrix}$, we get

$$\begin{aligned} \tilde{\varphi}_{i+} &= 1 - \mathcal{P}(z_i = 1 | Z_i = 1) = \frac{1 - \beta}{\alpha + \beta - 1} \left[\alpha \cdot \frac{\pi_-(\xi_i)}{\pi_+(\xi_i)} - (1 - \alpha) \right] \\ \tilde{\varphi}_{i-} &= 1 - \mathcal{P}(z_i = 0 | Z_i = 0) = \frac{1 - \alpha}{\alpha + \beta - 1} \left[\beta \cdot \frac{\pi_+(\xi_i)}{\pi_-(\xi_i)} - (1 - \beta) \right] \end{aligned} \quad (19)$$

226 These probabilities can be computed provided the odds of the test results $\frac{\pi_+(\xi_i)}{\pi_-(\xi_i)}$. We employ a parametric approach and
 227 approximate this odds by applying estimation tool M to the original data D ; i.e., the predicted binomial probability

228 $\hat{p}_1(\xi_i)$ obtained from D is used as a proxy of $\pi_+(\xi_i)$. Generally it is not assured that true and observed probabilities
 229 $p_1(\xi_i)$ and $\pi_+(\xi_i)$ have the same mechanistic structure captured by M ; however, when our concern is limited to
 230 the use of model-predicted probabilities to smooth the data D , we may expect for M to provide a sufficiently good
 231 approximation with realistic test sensitivity and specificity. The above framework can be regarded as a variant of
 232 parametric bootstrapping methods as MO datasets are generated from data D assuming a parametric model M . The
 233 whole bias correction procedure is presented in pseudocode (Algorithm 1); a sample R code is also available on GitHub
 234 (<https://github.com/akira-endo/TND-biascorrection/>).

Algorithm 1 Multiple imputation with parametric bootstrapping

Input: $D = \{Z_i, \xi_i\}_{i=1,2,\dots,S}$
 Fit model M to D to obtain a predictive model $\pi_+ = \hat{p}_1(\xi)$
for $j = 1, 2, \dots, J$ **do**
 for $i = 1, \dots, S$ **do**
 $\pi_+ \leftarrow \hat{p}_1(\xi_i)$ ▷ Predict the binomial probability π_+ by model M
 $\tilde{z}_i^j \leftarrow Z_i$ ▷ Copy Z_i , and then flip at a probability φ to impute \tilde{z}_i^j
 if $Z_i = 1$ **then**
 $\varphi \leftarrow \frac{1-\beta}{\alpha+\beta-1} \left[\alpha \cdot \frac{1-\pi_+}{\pi_+} - (1-\alpha) \right]$
 $u \leftarrow \text{Unif}(0, 1)$
 if $u < \varphi$ **then**
 $\tilde{z}_i^j \leftarrow 0$
 end if
 else
 $\varphi \leftarrow \frac{1-\alpha}{\alpha+\beta-1} \left[\beta \cdot \frac{\pi_+}{1-\pi_+} - (1-\beta) \right]$
 $u \leftarrow \text{Unif}(0, 1)$
 if $u < \varphi$ **then**
 $\tilde{z}_i^j \leftarrow 1$
 end if
 end if
 end for
 Fit model M to $\tilde{D}^j = \{\tilde{z}_i^j, \xi_i\}_{i=1,2,\dots,S}$ to estimate parameter ε^j
end for
Output: MO estimates $\{\varepsilon^j\}_{j=1,2,\dots,J}$

235 **(ii) EM algorithm**

236 Another possible approach is to use EM algorithm as proposed by Magder et al. [21], where $\tilde{\varphi}_{i\pm}$ can be estimated by
 237 iterations (see the supplementary document for details). However, in our simulation we found that the performance
 238 of EM algorithm was inferior to the other two alternatives (direct likelihood and parametric bootstrapping). The
 239 three methods all provided effectively identical distributions of estimates in most settings, but in some settings the
 240 EM algorithm produced extreme estimates ($\text{VE} < 0$ or > 1) slightly more often than the other two. We therefore
 241 recommend parametric bootstrapping as the first choice of bias correction method when the direct likelihood approach
 242 is inconvenient.

243 **2.3.4 Simulation of bias correction with parametric bootstrapping**

244 To assess the performance of this method, we used the same simulation framework as in the univariate analysis (Table
 245 1). In addition to vaccination history (denoted by ξ^1), we consider one categorical and one continuous covariate. Let us
 246 assume that ξ_2 represents the age group (categorical; 1: child, 0: adult) and ξ_3 the pre-infection antibody titre against
 247 TD (continuous). Suppose that the population ratio between children and adults is 1:2, and that ξ_3 is scaled so that
 248 it is standard normally distributed in the population. For simplicity, we assumed that all the covariates are mutually
 249 independent with regard to the distribution and effects (i.e., no association between covariates and no interaction effects).
 250 The relative risk of children was set to be 2 and 1.5 for TD and ND, respectively, and a unit increase in the antibody
 251 titre was assumed to halve the risk of TD (and not to affect the risk of ND). The mean total sample size λ was set
 252 to be 3,000, and 500 sets of simulation data were generated for each scenario. VE estimates were corrected by the
 253 parametric bootstrapping approach (the number of iterations $J = 100$) and were compared with the raw (uncorrected)
 254 VE estimates.

255 Figure 5 shows the distributions of estimates with and without bias correction in the multivariate setting. Our bias
 256 correction (parametric bootstrapping) provided unbiased estimates for all the scenarios considered. Overall, biases in
 257 the uncorrected estimates were larger than those in the univariate setting. In some scenarios, the standard error of the
 258 bias-corrected estimates was extremely wide. This was not because of the failure of bias correction, but because of the
 259 uncertainty already introduced before misclassification. The standard errors in those scenarios were very large
 260 even with perfect test sensitivity and specificity as can be seen in Figure 6. Larger sample size is required to yield
 261 accurate estimates in those settings, as the information loss due to misclassification will be added on top of the inherent
 262 uncertainty in the true data.

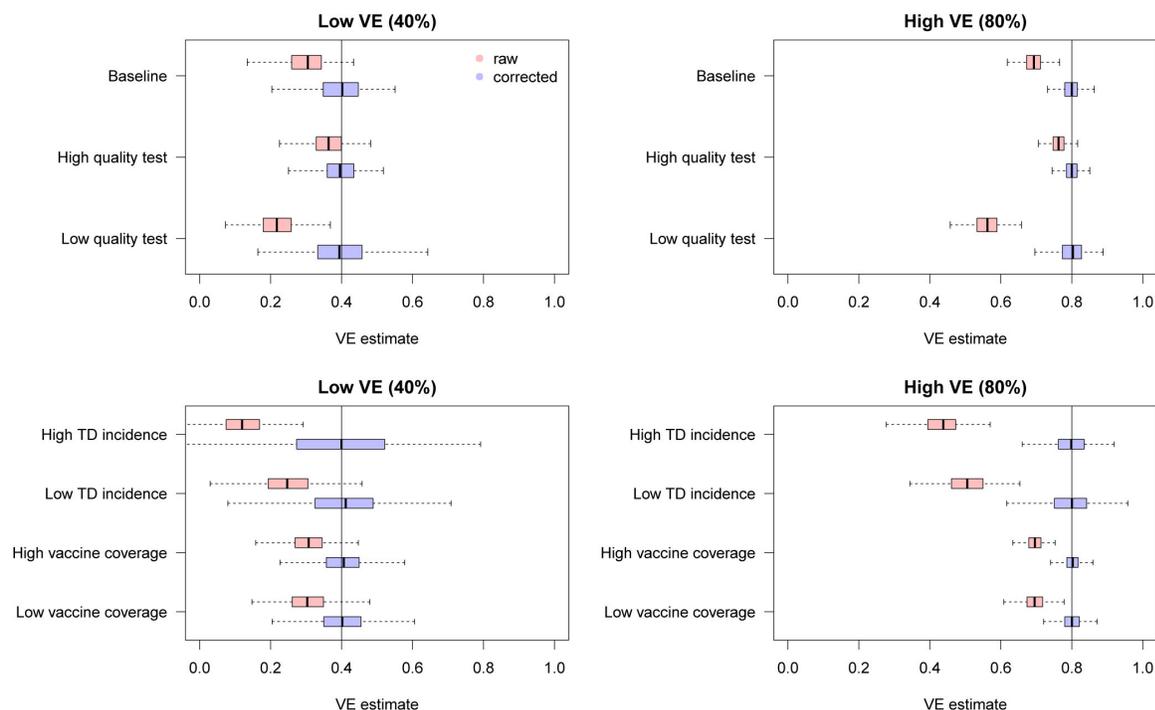


Figure 5: Bias correction for simulated data in the multivariate setting. The distributions of bias-corrected (blue) and uncorrected (red) VE estimates from 500 simulations are compared. Dotted lines denote median and black solid lines denote the true VE. The parametric bootstrapping bias correction method was used for bias correction.

263 2.3.5 Increased uncertainty introduced by misclassification

264 Although our bias correction methods provide unbiased VE estimates from potentially misclassified test results, the
 265 resulting uncertainty is larger than that which would be obtained from estimates using the true disease status. In Figure
 266 6, we compared bias-corrected estimates obtained from misclassified data (by the direct likelihood method) with those
 267 obtained from the true data (i.e., 100% sensitivity and specificity). Although both estimates are unbiased around the
 268 true value, the results from the misclassified data exhibit higher variability (by a factor of 1.1-3.0) due to the loss of
 269 information caused by misdiagnosis. Increased uncertainty due to misclassification should be carefully considered when
 270 one calculates the power of test-negative design studies. Overestimated test performance may not only underestimate
 271 the true VE but also lead to overconfidence.

272 2.3.6 The number of confounding variables

273 We investigated how the bias in uncorrected VE estimates can be affected by the number of confounding variables. In
 274 addition to the vaccine history ξ^1 , we added a set of categorical/continuous confounding variables to the model and
 275 assessed the degree of bias caused by misclassification. The characteristics of the variables were inherited from those in
 276 Section 2.3.4: categorical variable "age" and continuous variable "pre-infection antibody titre". That is, individuals
 277 were assigned multiple covariates (e.g., "categorical variable A", "categorical variable B", ..., "continuous variable

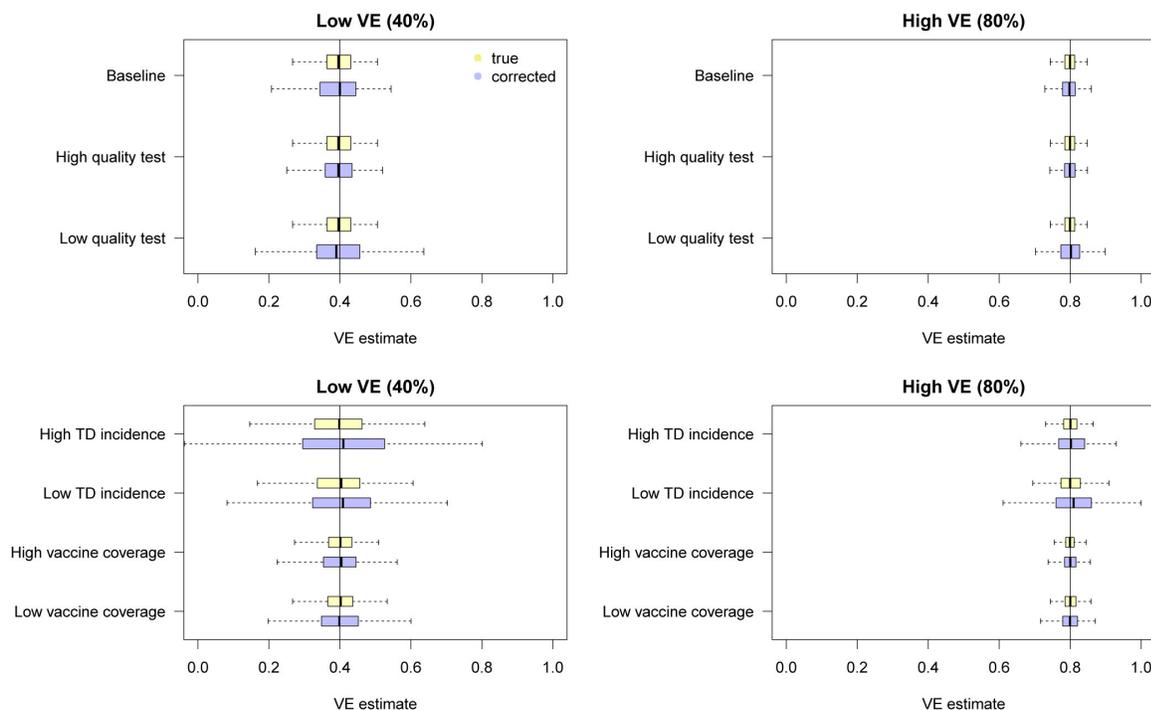


Figure 6: Uncertainty in VE estimates obtained from the true/misclassified datasets in the multivariate setting. The distributions of VE estimates from the simulated true (yellow) and misclassified (light blue) data are shown. The direct likelihood method was employed to correct biases in the misclassified data.

278 A", "continuous variable B", ...) whose distribution and effect were identical to "age" (for categorical variables) and
 279 "antibody titre" (for continuous variables) in Section 2.3.4. No interaction between covariates was assumed. The
 280 covariate set in Section 2.3.4 being baseline (the number of covariates: (vaccine, categorical, continuous) = (1, 1, 1)),
 281 we employed two more scenarios with a larger number of covariates: (1, 3, 3) and (1, 5, 5).

282 The simulation results are presented in Figure 7. Overall, additional confounding variables led to more severe bias in
 283 the uncorrected VE estimates towards underestimation. As shown in Figure 2A, the degree of bias is strongly affected
 284 by the case ratio: the ratio between the risk of TD and ND. More confounding variables in a population result in higher
 285 heterogeneity in individuals' risk of TD and ND. This may account for the association between the degree of bias and
 286 the number of confounding variables; more individuals in a highly heterogeneous population may fall in the outer range
 287 of the case ratio shown in Figure 2A, substantially contributing to the misclassification bias.

288 3 Discussion

289 Misclassification caused by imperfect diagnostic tests can potentially lead to substantial biases in TND studies. By
 290 considering the processes involved in VE estimation, we have characterised the degree of bias potentially caused
 291 by diagnostic misclassification in different parameter settings, finding that VE can be noticeably underestimated,
 292 particularly when the ratio between TD and ND cases in the study data is unbalanced. To address this potential bias, we
 293 developed multiple bias correction methods that address test misclassification and provide unbiased VE estimates in
 294 both univariate and multivariate settings. When the test sensitivity and specificity are known or assumed, those values
 295 can be used to restore the true VE estimate by a relatively simple statistical procedure. Using simulations, we showed
 296 that our methods could successfully eliminate the bias in VE estimates obtained from misclassified data, although some
 297 uncertainty was introduced as a result of the information loss.

298 We believe that our methods could therefore enable researchers to address possible misclassification in their data
 299 and report unbiased VE estimates even when imperfect tests had to be used. Such methods could also help in the
 300 scaling up of TND studies, as tests with limited performance are usually inexpensive and logistically convenient. Even

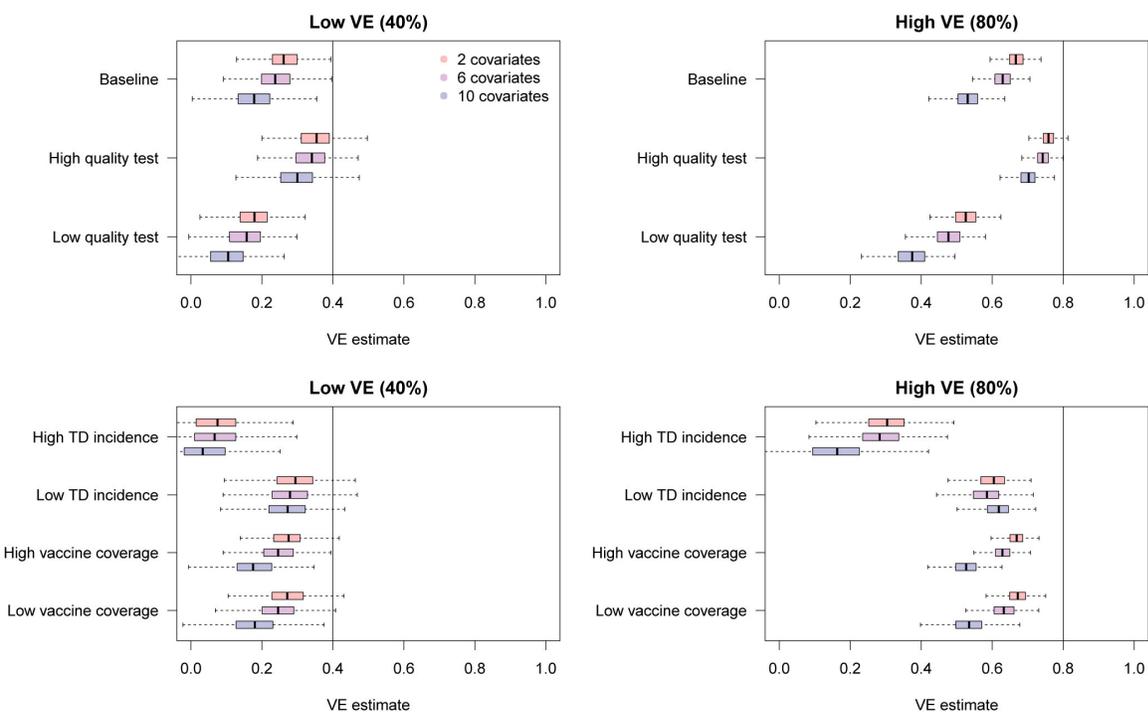


Figure 7: Bias in raw VE estimates from simulated data in the presence of different numbers of confounding variables. The distributions in red, purple and blue correspond to uncorrected VE estimates in the presence of 2, 6 and 10 confounding variables in addition to the vaccination history.

301 when high-performance diagnostic tests including PCR techniques are available, the risk of misdiagnosis may not be
 302 negligible in certain settings, and our methods could further be used to perform sensitivity analysis to properly address
 303 the possibility of bias in such cases.

304 Although TND is a relatively new study design, first appearing in a publication in 2005 [22], it has gained broad
 305 popularity and is becoming a standard approach in VE studies. TND is believed to minimise the bias caused by different
 306 health seeking behaviour of individuals, but one of the largest factors that have contributed to its widespread use is the
 307 fact that data collection can be completed within clinical setups [1]. Whereas cohort or case-control studies usually
 308 require additional efforts including follow up or recruitment of non-patients, TND studies only involve patients visiting
 309 healthcare facilities with suspects of certain diseases and thus routinely collected clinical data can be easily adapted for
 310 analysis. For diseases of which suspected patients routinely undergo lab tests, TND may be one of the most convenient
 311 options to generate epidemiological insights into the effect of specific prevention/treatment. VE studies of influenza, for
 312 which TND is most frequently used, often use PCR as a diagnostic tool for better data quality [20]. However, such
 313 studies usually involve intensive effort and cost, and thus may only be feasible by large-scale research bodies. Our
 314 bias correction methods may open a possibility of wider use of clinical data, which could potentially provide rich
 315 epidemiological insights, especially in settings where rapid tests are routinely used for diagnosis. For example, rapid
 316 influenza diagnostic tests are routinely used for inpatient clinics and hospitals in Japan, and such clinical data have
 317 facilitated a number of TND studies [23, 24, 25, 26, 27, 28]. Such studies based on rapid tests could benefit from
 318 our methods, as it would provide strong support for the validity of their estimates. Our methods may also be useful
 319 in resource limited settings or for diseases without high-performance diagnostic tools. Even in resourceful settings
 320 where high-performance tests are available, the slight possibility of misclassification might not always be neglected.
 321 Although PCR tests are currently used as a gold-standard for influenza diagnosis, their sensitivity and specificity may
 322 not be exact 100%; especially, the sensitivity of the test depends not only on microbiological technique but also on the
 323 quality of swab samples from patients. In addition, it is suggested that the sensitivity of PCR tests may change during
 324 the time course of infection [29] and be sufficiently high only during a limited time window. Our simulation study
 325 also indicated that a high heterogeneity in individual characteristics in study samples might increase the bias in the

326 VE estimate. Our methods could enable researchers to implement sensitivity analysis by assuming the possible test
327 sensitivity and specificity in such cases.

328 Our bias correction methods are also intended to be reasonably straightforward for researchers to introduce. Existing
329 estimation tools including software libraries and packages are often used in epidemiological analyses, and most of them
330 have specific requirements for data input and output. Modifying the procedure employed by such tools in a way which
331 is unexpected by the authors is usually impossible or requires advanced technical skills including reprogramming of
332 the underlying algorithms. Incorporating the MO approach, our parametric bootstrapping bias correction method only
333 involves data manipulation and does not require modification of the estimation algorithm. The only inputs required to
334 produce MO datasets are the assumed test sensitivity and specificity (α, β) and the model-predicted binomial probability
335 for the test result (π_+) for each individual. Once multiple sets of data are randomly generated, any type of analysis can
336 be performed as long as they produce numerical estimates to be summarised over the MO datasets. Of particular note is
337 that our methods for multivariate analysis (including the direct likelihood method) allow stratification of sensitivity and
338 specificity among individuals. Therefore, the users can employ more complex misclassification mechanisms including
339 time-varying test performance or test performance affected by individual characteristics. Datasets with a mixture of
340 different diagnostic tools [3, 30] can also be handled by applying different values for each test.

341 There are some limitations to our study. We only focused on misclassification of diagnosis (i.e., misclassified outcomes)
342 and did not consider potential misclassification of covariates (e.g., vaccine history and other confounding variables),
343 which is another important type of misclassification in TND studies [10]. Further, it is generally not easy to plausibly
344 estimate the sensitivity and specificity for measurement of covariates (e.g. recall bias), which must be known or assumed
345 to implement bias correction. However, if reliable estimates are available, an extension of our approach may yield
346 bias-corrected VE estimates in the presence of covariate misclassification. Such consideration remains to be discussed
347 in future work. Moreover, to keep our focus only on diagnostic misclassification, our methods rested on the assumption
348 that other sources of bias in TND studies are nonexistent or properly addressed. Potential sources of bias in TND
349 studies have been discussed elsewhere [13, 31], and the researchers conducting TND studies need to carefully consider
350 the possibility of such biases in addition to the diagnostic misclassification. Lastly, it must be noted that our methods
351 depend on the assumed test sensitivity and specificity, and that misspecifying those values can result in an improper
352 correction. The sensitivity and specificity of tests are usually reported by manufacturers in a comparison of the test
353 results with gold-standard tests; however, when such gold-standard tests themselves are not fully reliable or when no
354 available test has satisfactory performance to be regarded as gold-standard, specifying sensitivity and specificity of
355 a test is in principle impossible. Further, test performances reported by manufacturers might lack sufficient sample
356 size or might not be identical to those in the actual study settings. Use of composite reference standards [32, 33] or
357 external/internal validation approaches [34] may help overcome these problems.

358 Although the presence of imperfect diagnosis limits the quality of clinical data, data with such uncertainty can still hold
359 useful information, and this information can be transformed into useful insights by appropriate statistical processing.
360 Our bias correction methods were developed primarily for TND studies, but a similar approach could be applied to
361 broader classes of estimation problems with misclassification. The value of routinely collected data in healthcare
362 settings has become widely recognised with the advancement of data infrastructure, and we believe our methods could
363 help support the effective use of such data.

364 4 Conclusion

365 Bias correction methods for the test-negative design studies were developed to address potential misclassification bias
366 due to imperfect tests.

367 Appendix

368 Maximum likelihood estimates and confidence intervals in the univariate setting

369 Expanding Equation (4) in the main text, we get

$$\begin{aligned} \mathcal{L}(\gamma, \delta, \lambda_V, \lambda_U; D) &= \frac{[\alpha\gamma\delta + (1 - \beta)]^{X_V} [(1 - \alpha)\gamma\delta + \beta]^{Y_V} [\alpha\delta + (1 - \beta)]^{X_U} [(1 - \alpha)\delta + \beta]^{Y_U} \lambda_V^{S_V} \lambda_U^{S_U}}{(1 + \delta)^{S_V} (1 + \delta)^{S_U} X_V! Y_V! X_U! Y_U! \exp\left(\frac{1 + \gamma\delta}{1 + \delta} \lambda_V\right) \exp(\lambda_U)}, \end{aligned} \quad (20)$$

370 where $S_V = X_V + Y_V$ and $S_U = X_U + Y_U$.

371 For mathematical convenience, we change the variable λ_V to $\lambda'_V = \frac{1+\gamma\delta}{1+\delta}\lambda_V$. Let $l = \log \mathcal{L}(\gamma, \delta, \lambda'_V, \lambda_U; X)$. Partial
 372 derivatives of l are

$$\begin{aligned}\frac{\partial l}{\partial \gamma} &= \frac{\alpha\delta X_V}{\alpha\gamma\delta + (1-\beta)} + \frac{(1-\alpha)\delta Y_V}{(1-\alpha)\gamma\delta + \beta} - \frac{\delta S_V}{1+\gamma\delta} \\ \frac{\partial l}{\partial \delta} &= \frac{\alpha\gamma X_V}{\alpha\gamma\delta + (1-\beta)} + \frac{(1-\alpha)\gamma Y_V}{(1-\alpha)\gamma\delta + \beta} + \frac{\alpha X_U}{\alpha\delta + (1-\beta)} + \frac{(1-\alpha)Y_U}{(1-\alpha)\delta + \beta} - \frac{\gamma S_V}{1+\gamma\delta} - \frac{S_U}{1+\delta} \\ \frac{\partial l}{\partial \lambda'_V} &= \frac{S_V}{\lambda'_V} - 1 \\ \frac{\partial l}{\partial \lambda_U} &= \frac{S_U}{\lambda_U} - 1\end{aligned}\quad (21)$$

373 Equation (21) gives the maximum likelihood estimates:

$$\begin{aligned}\gamma^* &= \frac{X_V - \frac{1-\beta}{\beta}Y_V}{Y_V - \frac{1-\alpha}{\alpha}X_V} \cdot \frac{Y_U - \frac{1-\alpha}{\alpha}X_U}{X_U - \frac{1-\beta}{\beta}Y_U} \\ \delta^* &= \frac{\beta}{\alpha} \cdot \frac{X_U - \frac{1-\beta}{\beta}Y_U}{Y_U - \frac{1-\alpha}{\alpha}X_U} \\ \lambda'^*_V &= S_V \\ \lambda^*_U &= S_U\end{aligned}\quad (22)$$

374 The confidence intervals for parameters can be constructed using the Fisher's information matrix from Equation (21).
 375 λ'_V and λ_U are independent from other parameters and

$$\text{Var}(\lambda'_V) = -\frac{\partial^2 l}{\partial \lambda'^2_V} = \frac{S_V}{\lambda'^2_V} - \frac{\partial^2 l}{\partial \lambda^2_U} = \frac{S_U}{\lambda^2_U}\quad (23)$$

376 We log-transform γ and δ for mathematical convenience. Noting that $\frac{\partial y}{\partial(\log x)} = x \frac{\partial y}{\partial x}$, we get

$$\begin{aligned}-\frac{\partial^2 l}{\partial(\log(\gamma))^2} &= \frac{\gamma\delta}{(1+\gamma\delta)^2}S_V - \frac{\alpha(1-\beta)\gamma\delta}{[\alpha\gamma\delta + (1-\beta)]^2}X_V - \frac{(1-\alpha)\beta\gamma\delta}{[(1-\alpha)\gamma\delta + \beta]^2}Y_V \\ -\frac{\partial^2 l}{\partial \log \gamma \partial \log \delta} &= \frac{\gamma\delta}{(1+\gamma\delta)^2}S_V - \frac{\alpha(1-\beta)\gamma\delta}{[\alpha\gamma\delta + (1-\beta)]^2}X_V - \frac{(1-\alpha)\beta\gamma\delta}{[(1-\alpha)\gamma\delta + \beta]^2}Y_V \\ -\frac{\partial^2 l}{\partial(\log \delta)^2} &= \frac{\gamma\delta}{(1+\gamma\delta)^2}S_V - \frac{\alpha(1-\beta)\gamma\delta}{[\alpha\gamma\delta + (1-\beta)]^2}X_V - \frac{(1-\alpha)\beta\gamma\delta}{[(1-\alpha)\gamma\delta + \beta]^2}Y_V \\ &\quad + \frac{\delta}{(1+\delta)^2}S_U - \frac{\alpha(1-\beta)\delta}{[\alpha\delta + (1-\beta)]^2}X_U - \frac{(1-\alpha)\beta\delta}{[(1-\alpha)\delta + \beta]^2}Y_U\end{aligned}\quad (24)$$

377 With the parameter values estimated in Eq. (22), we get the following information matrix

$$\begin{bmatrix} \frac{x_V y_V}{S_V} \left[1 - S_V \left(\frac{\alpha(1-\beta)}{X_V} + \frac{(1-\alpha)\beta}{Y_V} \right) \right] & \frac{x_V y_V}{S_V} \left[1 - S_V \left(\frac{\alpha(1-\beta)}{X_V} + \frac{(1-\alpha)\beta}{Y_V} \right) \right] \\ \frac{x_V y_V}{S_V} \left[1 - S_V \left(\frac{\alpha(1-\beta)}{X_V} + \frac{(1-\alpha)\beta}{Y_V} \right) \right] & \frac{x_V y_V}{S_V} \left[1 - S_V \left(\frac{\alpha(1-\beta)}{X_V} + \frac{(1-\alpha)\beta}{Y_V} \right) \right] + \frac{x_U y_U}{S_U} \left[1 - S_U \left(\frac{\alpha(1-\beta)}{X_U} + \frac{(1-\alpha)\beta}{Y_U} \right) \right] \end{bmatrix}$$

378 where $x_\xi = \frac{1}{c}[\beta X_\xi - (1-\beta)Y_\xi]$ and $y_\xi = \frac{1}{c}[\alpha Y_\xi - (1-\alpha)X_\xi]$ are the true case counts (without misclassification)
 379 for $\xi = V, U$. Let $p_V = x_V/(x_V + y_V)$ and $p_U = x_U/(x_U + y_U)$ be the corresponding true binomial probabilities.

380 The inverse of the information matrix provides variance of estimates: in particular, for $\log \gamma$ we get

$$\begin{aligned} \text{Var}(\log \gamma^*) &= \frac{S_V}{x_V y_V} \cdot \frac{1}{\left[1 - \left(\frac{\alpha(1-\beta)}{\pi_V} + \frac{(1-\alpha)\beta}{1-\pi_V}\right)\right]} + \frac{S_U}{x_U y_U} \cdot \frac{1}{\left[1 - \left(\frac{\alpha(1-\beta)}{\pi_U} + \frac{(1-\alpha)\beta}{1-\pi_U}\right)\right]} \\ &= \frac{S_V}{x_V y_V} \cdot \frac{\pi_V(1-\pi_V)}{(1-\pi_V - (1-\alpha))(\pi_V - (1-\beta))} + \frac{S_U}{x_U y_U} \cdot \frac{\pi_U(1-\pi_U)}{(1-\pi_U - (1-\alpha))(\pi_U - (1-\beta))} \\ &= \frac{c^2}{S_V} \frac{\pi_V(1-\pi_V)}{(1-\pi_V - (1-\alpha))^2(\pi_V - (1-\beta))^2} + \frac{c^2}{S_U} \frac{\pi_U(1-\pi_U)}{(1-\pi_U - (1-\alpha))^2(\pi_U - (1-\beta))^2}. \end{aligned} \quad (25)$$

381 We can relate this to the true standard error that would be obtained with perfect tests,

$$\text{SD}(\log \gamma_{\text{true}}) = \sqrt{\frac{1}{S_V p_V(1-p_V)} + \frac{1}{S_U p_U(1-p_U)}} = \sqrt{\frac{\sigma_V^2}{S_V} + \frac{\sigma_U^2}{S_U}}, \quad (26)$$

382 or to the observed standard error (without correction),

$$\text{SD}(\log \gamma_{\text{raw}}) = \sqrt{\frac{1}{S_V \pi_V(1-\pi_V)} + \frac{1}{S_U \pi_U(1-\pi_U)}} = \sqrt{\frac{\Sigma_V^2}{S_V} + \frac{\Sigma_U^2}{S_U}}, \quad (27)$$

383 where $\sigma_V = [p_V(1-p_V)]^{-1/2}$ and $\sigma_U = [p_U(1-p_U)]^{-1/2}$ are the components of the true standard error and
384 $\Sigma_V = [\pi_V(1-\pi_V)]^{-1/2}$ and $\Sigma_U = [\pi_U(1-\pi_U)]^{-1/2}$ are those of uncorrected standard error. We get

$$\begin{aligned} \sigma^* = \text{SD}(\log(\gamma^*)) &= \sqrt{\frac{\sigma_V^2}{S_V} \cdot \frac{1}{\left(1 - \frac{1-\alpha}{1-\pi_V}\right)\left(1 - \frac{1-\beta}{\pi_V}\right)} + \frac{\sigma_U^2}{S_U} \cdot \frac{1}{\left(1 - \frac{1-\alpha}{1-\pi_U}\right)\left(1 - \frac{1-\beta}{\pi_U}\right)}} \\ &= \frac{1}{c} \sqrt{\frac{\Sigma_V^2}{S_V} \cdot \left(\frac{\pi_V(1-\pi_V)}{p_V(1-p_V)}\right)^2 + \frac{\Sigma_U^2}{S_U} \cdot \left(\frac{\pi_U(1-\pi_U)}{p_U(1-p_U)}\right)^2}. \end{aligned} \quad (28)$$

385 This equation indicates that the confidence intervals diverge when the true outcome is bipolarised ($p_V, p_U \simeq 0$ or 1).

386 Declarations

387 This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit
388 sectors. Authors declare that they have no competing interests.

389 References

- 390 [1] G. De Serres, D. M. Skowronski, X. W. Wu, and C. S. Ambrose. The test-negative design: Validity, accuracy and
391 precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical
392 trials. *Eurosurveillance*, 2013.
- 393 [2] Wakaba Fukushima and Yoshio Hirota. Basic principles of test-negative design in evaluating influenza vaccine
394 effectiveness. *Vaccine*, 2017.
- 395 [3] Motoi Suzuki, Bhim Gopal Dhoubhadel, Tomoko Ishifuji, Michio Yasunami, Makito Yaegashi, Norichika Asoh,
396 Masayuki Ishida, Sugihiro Hamaguchi, Masahiro Aoshima, Koya Ariyoshi, and Konosuke Morimoto. Serotype-
397 specific effectiveness of 23-valent pneumococcal polysaccharide vaccine against pneumococcal pneumonia in
398 adults aged 65 years or older: a multicentre, prospective, test-negative design study. *The Lancet Infectious
399 Diseases*, 17(3):313–321, mar 2017.
- 400 [4] Rosa Prato, Francesca Fortunato, Maria Giovanna Cappelli, Maria Chironna, and Domenico Martinelli. Effective-
401 ness of the 13-valent pneumococcal conjugate vaccine against adult pneumonia in italy: a case-control study in a
402 2-year prospective cohort. *BMJ Open*, 8(3), 2018.

- 403 [5] Kaoru Araki, Megumi Hara, Takeshi Tsugawa, Chisato Shimanoe, Yuichiro Nishida, Muneaki Matsuo, and
404 Keitaro Tanaka. Effectiveness of monovalent and pentavalent rotavirus vaccines in Japanese children. *Vaccine*,
405 36(34):5187 – 5193, 2018.
- 406 [6] Anna Lena Lopez, Jeda Veronica Daag, Joel Esparagoza, Joseph Bonifacio, Kimberley Fox, Batmunkh Nyambat,
407 Umesh D Parashar, Maria Joyce Ducusin, and Jacqueline E Tate. Effectiveness of monovalent rotavirus vaccine in
408 the Philippines. *Scientific reports*, 8(1):14291, sep 2018.
- 409 [7] K Muhsen, E Anis, U Rubinstein, E Kassem, S Goren, L M Shulman, M Ephros, and D Cohen. Effectiveness
410 of rotavirus pentavalent vaccine under a universal immunization programme in Israel, 2011–2013;2015: a
411 case–control study. *Clinical Microbiology and Infection*, 24(1):53–59, jan 2018.
- 412 [8] Evan W. Orenstein, Gaston De Serres, Michael J. Haber, David K. Shay, Carolyn B. Bridges, Paul Gargiullo,
413 and Walter A. Orenstein. Methodologic issues regarding the use of three observational study designs to assess
414 influenza vaccine effectiveness. *International Journal of Epidemiology*, 2007.
- 415 [9] Michael L. Jackson and Kenneth J. Rothman. Effects of imperfect test sensitivity and specificity on observational
416 studies of influenza vaccine effectiveness. *Vaccine*, 2015.
- 417 [10] Tom De Smedt, Elizabeth Merrall, Denis Macina, Silvia Perez-Vilar, Nick Andrews, and Kaatje Bollaerts. Bias
418 due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness.
419 *PLOS ONE*, 2018.
- 420 [11] Sander Greenland. Basic methods for sensitivity analysis of biases, 1996.
- 421 [12] Matthew Blackwell, James Honaker, and Gary King. A Unified Approach to Measurement Error and Missing
422 Data: Overview and Applications. *Sociological Methods and Research*, 2017.
- 423 [13] M. Haber, Q. An, I. M. Foppa, D. K. Shay, J. M. Ferdinands, and W. A. Orenstein. A probability model
424 for evaluating the bias and precision of influenza vaccine effectiveness estimates from case-control studies.
425 *Epidemiology and Infection*, 143(7):1417–1426, 2015.
- 426 [14] Benjamin J. Cowling and Hiroshi Nishiura. Virus Interference and Estimates of Influenza Vaccine Effectiveness
427 from Test-Negative Studies. *Epidemiology*, 2012.
- 428 [15] Increased risk of noninfluenza respiratory virus infections associated with receipt of inactivated influenza vaccine.
429 *Clinical Infectious Diseases*, 2012.
- 430 [16] Maria E. Sundaram, David L. McClure, Jeffrey J. Vanwormer, Thomas C. Friedrich, Jennifer K. Meece, and
431 Edward A. Belongia. Influenza vaccination is not associated with detection of noninfluenza respiratory viruses in
432 seasonal studies of influenza vaccine effectiveness. *Clinical Infectious Diseases*, 2013.
- 433 [17] M. Suzuki, A. Camacho, and K. Ariyoshi. Potential effect of virus interference on influenza vaccine effectiveness
434 estimates in test-negative designs. *Epidemiology and Infection*, 2014.
- 435 [18] V. K. Leung, B. J. Cowling, S. Feng, and Sheena G. Sullivan. Concordance of interim and final estimates of
436 influenza vaccine effectiveness: A systematic review. 2016.
- 437 [19] Barnaby Young, Sapna Sadarangani, Lili Jiang, Annelies Wilder-Smith, and Mark I.Cheng Chen. Duration of
438 influenza vaccine effectiveness: A systematic review, meta-analysis, and meta-regression of test- negative design
439 case-control studies. *Journal of Infectious Diseases*, 2018.
- 440 [20] Effectiveness of seasonal influenza vaccine in community-dwelling elderly people: A meta-analysis of test-negative
441 design case-control studies. *The Lancet Infectious Diseases*, 2014.
- 442 [21] Laurence S. Magder and James P. Hughes. Logistic regression when the outcome is measured with uncertainty.
443 *American Journal of Epidemiology*, 146(2):195–203, 1997.
- 444 [22] Effectiveness of vaccine against medical consultation due to laboratory-confirmed influenza: results from a
445 sentinel physician pilot project in British Columbia, 2004–2005. *Canada Communicable Disease Report = Relevé*
446 *des Maladies Transmissibles au Canada*, 2005.
- 447 [23] Yuki Seki, Hiroka Oonishi, Akira Onose, and Norio Sugaya. [Effectiveness of Influenza Vaccine in Adults Using
448 A Test-negative, Case-control Design -2013/2014 and 2014/2015 Seasons-](Japanese). *Kansenshogaku zasshi*.
449 *The Journal of the Japanese Association for Infectious Diseases*, 2016.
- 450 [24] Yuki Seki, Akira Onose, and Norio Sugaya. Influenza vaccine effectiveness in adults based on the rapid influenza
451 diagnostic test results, during the 2015/16 season. *Journal of Infection and Chemotherapy*, 2017.
- 452 [25] Nobuo Saito, Kazuhiro Komori, Motoi Suzuki, Kounosuke Morimoto, Takayuki Kishikawa, Takahiro Yasaka, and
453 Koya Ariyoshi. Negative impact of prior influenza vaccination on current influenza vaccination among people
454 infected and not infected in prior season: A test-negative case-control study in Japan. *Vaccine*, 2017.

- 455 [26] Masayoshi Shinjoh, Norio Sugaya, Yoshio Yamaguchi, Noriko Iibuchi, Isamu Kamimaki, Anna Goto, Hisato
456 Kobayashi, Yasuaki Kobayashi, Meiwa Shibata, Satoshi Tamaoka, Yuji Nakata, Atsushi Narabayashi, Mitsuhiro
457 Nishida, Yasuhiro Hirano, Takeshi Munenaga, Kumiko Morita, Keiko Mitamura, and Takao Takahashi. Inactivated
458 influenza vaccine effectiveness and an analysis of repeated vaccination for children during the 2016/17 season.
459 *Vaccine*, 2018.
- 460 [27] Soichiro Ando. Effectiveness of quadrivalent influenza vaccine based on the test-negative control study in children
461 during the 2016–2017 season. *Journal of Infection and Chemotherapy*, 2018.
- 462 [28] Norio Sugaya, Masayoshi Shinjoh, Yuji Nakata, Kenichiro Tsunematsu, Yoshio Yamaguchi, Osamu Komiyama,
463 Hiroki Takahashi, Keiko Mitamura, Atsushi Narabayashi, and Takao Takahashi. Three-season effectiveness of
464 inactivated influenza vaccine in preventing influenza illness and hospitalization in children in Japan, 2013-2016.
465 *Vaccine*, 2018.
- 466 [29] Sheena G. Sullivan, Shuo Feng, and Benjamin J. Cowling. Potential of the test-negative design for measuring
467 influenza vaccine effectiveness: A systematic review. *Expert Review of Vaccines*, 2014.
- 468 [30] Comparison of two control groups for estimation of oral cholera vaccine effectiveness using a case-control study
469 design. *Vaccine*, 2017.
- 470 [31] Joseph A. Lewnard, Christine Tedijanto, Benjamin J. Cowling, and Marc Lipsitch. Measurement of vaccine direct
471 effects under the test-negative design. *American Journal of Epidemiology*, 2018.
- 472 [32] Direk Limmathurotsakul, Elizabeth L. Turner, Vanaporn Wuthiekanun, Janjira Thaipadungpanit, Yupin Suputta-
473 mongkol, Wirongrong Chierakul, Lee D. Smythe, Nicholas P.J. Day, Ben Cooper, and Sharon J. Peacock. Fool's
474 gold: Why imperfect reference tests are undermining the evaluation of novel diagnostics: A reevaluation of 5
475 diagnostic tests for leptospirosis. *Clinical Infectious Diseases*, 2012.
- 476 [33] Christiana A. Naaktgeboren, Loes C.M. Bertens, Maarten van Smeden, Joris A.H. de Groot, Karel G.M. Moons,
477 and Johannes B. Reitsma. Value of composite reference standards in diagnostic research. *BMJ (Clinical research*
478 *ed.)*, 2013.
- 479 [34] Suzan R. Kahn, Elham Rahme, Lisa M. Lix, Mark Burman, Geneviève Lefebvre, Jiayi Ni, Kaberi Dasgupta,
480 Yves Laflamme, Greg Berry, Ronald Dimentberg, Denis Talbot, and Alain Cirkovic. Comparing external and
481 internal validation methods in correcting outcome misclassification bias in logistic regression: A simulation study
482 and application to the case of postsurgical venous thromboembolism following total hip and knee arthroplasty.
483 *Pharmacoepidemiology and Drug Safety*, 2018.