American Journal of Epidemiology, 2025, 194, 2570–2579
https://doi.org/10.1093/aje/kwae423
Advance access publication date October 29, 2024
Original Contribution

# Addressing bias due to measurement error of an outcome with unknown sensitivity in database epidemiologic studies. A contribution from the ConcePTION project

Giorgio Limoncella*[,1], Leonardo Grilli[1], Emanuela Dreassi[1], Carla Rampichini[1], Robert Platt[2], Rosa Gini[3]

[1]Department of Statistics, Informatics and Applications "Giuseppe Parenti", University of Florence, Florence 50134, Italy
[2]Departments of Pediatrics and of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec H3A 0G3, Canada
[3]ARS (Agenzia Regionale di Sanità della Toscana), Florence 50141, Italy

*Corresponding author: Giorgio Limoncella, Università degli Studi di Firenze, Viale Giovanni Battista Morgagni, 59, 50134 Firenze FI (giorgio.limoncella@unifi.it)

## Abstract

In epidemiologic database studies, the occurrence of an event is measured with error through an indicator whose specificity is often maximized, at the expense of sensitivity. However, if the indicator has low sensitivity, measures of occurrence are underestimated. In association studies, risk difference is biased, and risk ratio may be biased as well, in either direction, if the sensitivity is differential across exposure groups. In this work, we show that if an auxiliary screening indicator can be defined to complement the main indicator, estimates of the positive predictive value of both indicators provide tools to estimate the sensitivity of the primary indicator or a lower bound thereof. This mitigates bias in estimating the number of cases, prevalence, cumulative incidence, rate (particularly when the event is rare), and, in association studies, risk ratio and risk difference. They also allow testing for nondifferential sensitivity. Although direct estimation of sensitivity is often infeasible, this novel methodology improves evidence based on data obtained from reuse of existing databases, which may prove critical for regulatory and public health decisions.

**Key words**: measurement error; validity indices; bootstrap test; nondifferential sensitivity; adjustment for misclassification.

## Introduction

Electronic health care data sources are a key resource in assessing the occurrence of events in human populations, playing a major role in public health and regulatory decision-making. During the recent coronavirus crisis, calculation of background rates of events that may have later occurred as adverse reactions to the new vaccines were requested by the European Medicines Agency ahead of the vaccination campaign and, in March 2021, were used to rapidly assess potential causality of safety signals emerging from spontaneous reports.[1,2] In pregnancy studies, the occurrence of adverse maternal or infant outcomes in the general population and in populations exposed to medications can be used to provide signals on the safety of use by pregnant women, because clinical trials in this population are rarely conducted.[3] Information on the prevalence of conditions such as hypertensive disorders during pregnancy is also important for pregnancy studies.[4]

In database studies, the occurrence of the event of interest is measured using an indicator. Due to the nature of such databases, the indicator may be affected by errors: false positives are cases that are detected by the indicator, but are not true cases; false negatives are true cases that are missed by the indicator. The degree of misclassification of an indicator is measured by its validity indices,[5] including sensitivity (SE), which is the proportion of true positives among true cases in the population, specificity (SP), which is the proportion of true negatives among noncases in the population, and positive predictive value (PPV), which is the proportion of true positives among cases detected by the indicator. Guidelines in the conduct and reporting of database studies recommend to estimate validity indices,[6] which would allow adjustment of observed rates using a straightforward correction (PPV/SE).[7]

However, validation is often not feasible due to cost, time constraints, or ethical reasons.[8] When validation studies are conducted, they often estimate only the PPV. Indeed, estimation of the PPV requires extracting a sample of cases positive for the indicator and their assessment through comparison with a gold standard. On the contrary, the true outcome is not observed, and it is often a rare event, thus estimating the sensitivity requires large validation samples to get enough true positive cases.[9-15] Therefore, the rate of occurrence is overestimated if SE > PPV, or underestimated if adjusted only with the PPV or if PPV > SE.

A second problem is when the objective of the database study is to assess a causal relationship between an exposure and an outcome. In this case, misclassification of the outcome or the

**Received:** January 31, 2024. **Accepted:** August 27, 2024

exposure may introduce bias both in the risk ratio and in the risk difference.[16] Each validity index can be defined with respect to the exposure groups, and the SE or SP of the indicator of the outcome is said to be nondifferential if it does not depend on the exposure group. A commonly accepted idea in epidemiology is that nondifferential misclassification of the exposure skews the expected value of an estimator toward the null. Another established principle regarding misclassification of the outcome is that if the SE is nondifferential and the PPV is 1, then the RR is unbiased (for this reason, guidelines recommend choosing indicators that minimize false positives, so the PPV approaches 1). However, the mentioned results may not hold exactly, because even if the data generation process has nondifferential errors, observed errors can still be differential, especially in small populations.[17] Furthermore, it has been demonstrated that when misclassification of exposure is not independent of misclassification of the outcome, even a small error can significantly bias the results in both directions, especially in scenarios with low disease prevalence.[18,19] In fact, even if we may assume that risk ratio is not biased, risk difference is prone to bias irrespective of whether SE is differential.

In this article, we build on a strategy suggested by Lanes and Beachler[20] to correct outcome misclassification, while we assume the exposure to be recorded without error. Causes for misclassification of exposure are classified in a recent study[21] and include purchase over-the-counter, medicines prescribed by a different prescriber or to a different patient, and medicines prescribed but not collected or not ingested/administered. The assumption of perfect measurement of exposure holds when such cases are less likely to happen, such as when a medicine is newly introduced in the market, with an uncertain risk profile, and/or is costly, because prescription, dispensing, and administration are expected to be more strictly controlled by the health care system that is originating the health care data.

Alongside a primary outcome indicator with a high PPV, we suggest searching for an auxiliary screening indicator aimed at capturing all (or nearly all) the cases so it has near-perfect SE, while still being specific enough to exclude many noncases.[20,22] Estimating the PPV of both indicators then becomes feasible, possibly stratified per exposure.

We consider the situation when a screening algorithm is available. We have an estimate of PPV for both indicators, and we are able to make at least 1 in a range of assumptions. We first show it is possible to estimate the SE of the primary indicator or a lower bound thereof. Furthermore, in the case when an exposure is involved in the study and the PPV of both indicators is estimated across exposure strata, we provide a hypothesis test to verify whether the SE of the main indicator is nondifferential. A simulation study is conducted to verify the performance of the test in multiple scenarios. Finally, we demonstrate how to use the SEs estimated from the PPVs to adjust, or to obtain bounds on, the number of cases, risks, risk ratios, and risk differences.

# Methods

## Notation

We denote by $Y$ the binary variable for the true value of the outcome of interest, with $A$ being an observed, possibly misclassified, binary variable that is the primary indicator for $Y$, and by $B$ being a screening indicator for the same outcome. Combining $A$ and $B$ yields further indicators, such as $A \cup B$, namely the indicator retrieving all cases retrieved by either $A$ or $B$, and $A \cap B$, which is the intersection indicator retrieving all cases retrieved by both $A$ and $B$.

We assume a framework with a finite target population with $N$ subjects. For an indicator $I$, we denote by $TP_I$, $FN_I$, and $FP_I$ the true positives, false negatives, and false positives, respectively, in the target population. Then, $N_Y = TP_I + FN_I$ is the number of subjects in the study population for whom the outcome has occurred and $N_I = TP_I + FP_I$ is the number of subjects identified by $I$.

We indicate with $\pi = N_Y/N$ the true, unobserved prevalence of Y; we indicate with $P_I = N_I/N$ the observed prevalence. Note that $\pi$ does not depend on the indicator.

The validity indices are defined for each indicator $I$ as follows: $SE_I = TP_I/(TP_I + FN_I)$; $SP_I = TN_I/(TN_I + FP_I)$; and $PPV_I = TP_I/(TP_I + FP_I)$.

We focus on the case of a binary exposure, denoting with $e$ the exposed and with $\bar{e}$ the nonexposed. We denote by $N^e$ and $N^{\bar{e}}$ the number of subjects in the study population who are in the exposed and nonexposed groups, respectively. Then, $SE_I^e$ is the SE for the exposed and $SE_I^{\bar{e}}$ is the SE for the nonexposed. Similarly, $SP_I^e$ is the SP for the exposed and $SP_I^{\bar{e}}$ is the SP for the nonexposed. An indicator $I$ has nondifferential SE if $SE_I^e = SE_I^{\bar{e}}$ and nondifferential SP if $SP_I^e = SP_I^{\bar{e}}$.

Lastly, we adopt the notation in which the "hat" symbol above a parameter (eg, $\hat{PPV}$) indicates we are referring to a sample statistic rather than the population value of the parameter.

## SE of indicators

### *Estimating the SE*

We first consider the situation where $B$ is a comprehensive screening indicator, namely, an indicator for which $SE_{A\cup B} = 1$, though it can include false positives. The following formula, derived from the interrelation between validity indices as described in Appendix S1 and proven in Appendix S2, shows how the SE of the primary indicator can be derived from observed prevalences and PPVs:

$$SE_A = 1 - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A + P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}} \quad (1)$$

As an example, consider a study on myocardial infarction under the assumption that all case patients who are not promptly admitted to a hospital die. In this case, $A$ is the indicator retrieving hospital admissions with a specific diagnostic code of infarction, whereas $B$ is the indicator retrieving hospitalization with unspecified codes of infarction or death for any cause.[20]

If $SE_{A\cup B}$ cannot be assumed to be 1, it is proven in Appendix S2 that the right-hand side of eqn (1) is an upper bound for the SE of $A$, namely,

$$SE_A \leq 1 - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A + P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}} \quad (2)$$

If no case is retrieved by both $A$ and $B$, namely $P_{A\cap B} = 0$, eqn (2) simplifies to

$$SE_A \leq \frac{P_A \times PPV_A}{P_A \times PPV_A + P_B \times PPV_B} \quad (3)$$

The approach summarized by eqns (1), (2), and (3) has the advantage of providing an estimate of the SE based on PPV; the drawback is that the estimator of the SE is based on the estimators of 3 PPVs, namely those of the primary indicator $A$, the screening indicator $B$, and their intersection $A \cap B$. This could imply a loss of accuracy due to conveying multiple sources of uncertainty. Equations (1), (2), and (3) can be written using the absolute number of

subjects identified by an indicator ($N_I$) in place of the observed prevalence ($P_I$). Thus, by way of illustration, eqn (3) becomes:

$$SE_A \leq \frac{N_A \times PPV_A}{N_A \times PPV_A + N_B \times PPV_B} \quad (4)$$

### *Example: Estimating an upper bound for the SE of an indicator of angioedema*

This example was presented at the 38th International Conference on Pharmacoepidemiology and Therapeutic Risk Management.[23] Angioedema is a local, circumscribed edema due to increased plasma leakage from capillaries into the deep layers of the skin and mucous membranes. The Entresto study is a postauthorization safety study requested by the European Medicines Agency as part of the Risk Management Plan of Entresto, a medicinal product indicated for the treatment of heart failure. One of the objectives of the study was to estimate the incidence and relative risk (RR) of angioedema, which is a suspected adverse reaction to the medicinal product.[24]

As part of the Entresto study, a validation study was conducted in the data source, accessed by ARS Toscana, to estimate the PPV and SE of an indicator *A* for angioedema (*International Classification of Diseases, Ninth Revision, Clinical Modification* [ICD-9-CM] code 995.1). A screening indicator *B* was identified, retrieving cases of hypersensitivity (ICD-9-CM codes 374.82, 376.33, 478.25, 478.6, 478.75, 508.8, 708.0, 708.1, 708.8, 708.9, 782.3995.0, 995.2, 995.27). In the study population, *A* retrieved $N_A = 34$ cases and *B* retrieved $N_B = 451$ cases, whereas no case was found by both *A* and *B*; hence, $N_{A\cap B} = \mathbf{0}$. All cases retrieved by *A* were validated, and 12 were found to be true cases; hence, $\hat{P}PV_A = \mathbf{35.3}\%$. A sample of 96 cases was validated from those retrieved by *B*, and 8 were found to be true cases; hence, $\hat{P}PV_B = \mathbf{8.3}\%$. Therefore, from eqn (4), $\hat{S}E_A \leq \mathbf{24.5}\%$.

Note that even though the PPV of *B* is much lower than the PPV of *A*, because its observed prevalence is much higher, the true cases among those retrieved by *B* may be a substantial share. This will be quantified in section "Number of cases and risk".

## A test for nondifferential SE

The nondifferentiality hypothesis states that the SE of the primary indicator is identical across exposure groups, namely

$$H_0 : SE_A^e = SE_A^{\bar{e}} \quad (5)$$

To carry out the test for the hypothesis of nondifferential SE, hypothesis (5) can be rewritten as

$$H_0 : \frac{SE_A^e}{SE_A^{\bar{e}}} - 1 = 0 \quad (6)$$

To derive the test statistic, we estimate the sensitivities of *A* in the exposure strata by exploiting a screening indicator *B*. Specifically, in Appendix S3, we prove that, under the assumption of nondifferentiality of the SE of $A \cup B$ (ie, $SE_{A\cup B}^e = SE_{A\cup B}^{\bar{e}}$), hypothesis (6) can be expressed in terms of observed prevalences (*P*) and PPV as follows:

$$H_0 : \frac{P_A^e PPV_A^e \left(P_A^{\bar{e}} PPV_A^{\bar{e}} + P_B^{\bar{e}} PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} PPV_{A\cap B}^{\bar{e}}\right)}{P_A^{\bar{e}} PPV_A^{\bar{e}} \left(P_A^e PPV_A^e + P_B^e PPV_B^e - P_{A\cap B}^e PPV_{A\cap B}^e\right)} - 1 = 0 \quad (7)$$

The estimate of the PPV, denoted as $\hat{P}PV$, is achieved by comparing each subject with a gold standard, which provides the true classification of the outcome. The validation study must be carefully designed to provide accurate estimates of the PPVs, and the sampling process must consider each exposure stratum and each indicator.

Once the required PPVs are estimated, given that the observed prevalences *P* are known, it is straightforward to calculate the value of test statistic. To carry out the test, it is required to determine the distribution of the test statistic under the null. This is not straightforward; therefore, we proceed with the bootstrap method[25] and reject the hypothesis if the 95% CI does not include 0.

## Simulation study

To assess the performance of the proposed nondifferentiality test, we conducted a simulation study over multiple scenarios in which the indicators' validity indices and the population characteristics were varied. We generated 2 binary vectors: 1 representing the true outcome of interest *Y* and the other representing the exposure groups $e$ and $\bar{e}$. The proportion of subjects exposed was fixed at 5%. We explored 3 scenarios for the true prevalence of the outcome ($Y = 1$) in the unexposed group: $\pi^{\bar{e}} = \mathbf{0.01}$ (rare), $\pi^{\bar{e}} = \mathbf{0.05}$, and $\pi^{\bar{e}} = \mathbf{0.1}$ (common). Moreover, we considered 3 scenarios for the RR: $RR = \pi^e/\pi^{\bar{e}} = 1$, 1.2, and 2.

Then, we generated 2 indicators of *Y*, denoted with *A* (primary) and *B* (screening), which are misclassified with specific error rates for each exposure group. $SE_{A\cup B}$ was fixed to 0.85 or 0.95. In each scenario, $SE_{A\cup B}$ remains equal in both exposure groups to comply with the assumption of nondifferentiality of the SE of $A \cup B$.

The values of the SEs are chosen to produce an SE ratio $SE_A^e/SE_A^{\bar{e}}$ in {0.6, 0.8, 1, 1.25, 1.67}, where values greater than 1 are reciprocals of those less than 1 (in fact, 1.25 = 1/0.6 and 1.67 = 1/0.8). In particular, when $SE_{A\cup B} = \mathbf{0.85}$, then $SE_A^{\bar{e}}$ was set at 0.5 and $SE_A^e$ varied in {0.3, 0.4, 0.5, 0.625, 0.835}; alternatively, when $SE_{A\cup B} = \mathbf{0.95}$, then $SE_A^e$ was fixed at 0.75 and $SE_A^{\bar{e}}$ varied in {0.45, 0.60, 0.75, 0.94} (in this case, the fifth value is missing because it is impossible to produce a SE ratio equal to 1.67).

To study the possible scenarios in real-life situations, we varied the intersection among *A* and *B*, considering 2 values for $SE_{A\cap B}$: 0 and 0.2. When $SE_{A\cap B}$ was equal to 0, $SP_{A\cap B}$ was set to 1 to create a nonintersection scenario in which no subject tests positive for both indicators. Nonintersection scenarios are plausible, as seen in the Entresto study described in Section "Example: estimating an upper bound for the sensitivity of an indicator of angioedema".

The SP was set to $1-\pi^{\bar{e}}/\mathbf{10}$ and $1-\pi^{\bar{e}}$ for *A* and *B*, respectively, in both exposure groups; this implies that the PPV has a maximum of 90% for *A* and 50% for *B*. We further assumed that $1 - SP_{A\cap B} = (\mathbf{1} - SP_A) \times (\mathbf{1} - SP_B)$, that is, the number of false positives of *A* is independent of the number of false positives of *B*.

For each scenario, we generated a finite population of size 1 million with the aforementioned characteristics. Then, repeating 1000 times, we drew a validation sample and carried out the test. We considered 3 different sizes of the validation sample: 250, 500, and 750.

We carried out random sampling across indicators and exposure. Specifically, we sampled 40% of the subjects from those with $A = 1$, 40% from $B = 1$, and 20% from $A = 1$ and $B = 1$ jointly. Each of these groups has 50% exposed and 50% unexposed. In eqn (7), the PPV was replaced by its estimate $\hat{P}PV$, given by the proportion of true positives ($Y = 1$) in the subset of subjects flagged by the indicator ($A = 1$ or $B = 1$).

The nondifferentiality test was conducted using the bootstrap method.[25] Specifically, 500 bootstrap samples were generated from the validation sample. For each bootstrap sample, the test statistic was calculated, and its empirical distribution was derived. The 95% CI was then computed, and the null hypothesis was rejected if the interval did not encompass zero. The power of the test is the proportion of times the test rejected the null hypothesis, denoted as $Pr_{rej}$. When the null hypothesis is true (SE ratio = 1), $Pr_{rej}$ is the empirical significance level of the test.

The simulation was conducted using R software. The code is freely available on GitHub at the following link: https://github.com/GiorgioLimoncella/NonDifferentialityTest.

The findings of the simulation study are summarized in Figure 1 and Figure 2, which show how the rejection rate $Pr_{rej}$ varies according to the true SE ratio in the case when $SE^{\bar{e}} = 0.5$ and $SE^{\bar{e}} = 0.75$, respectively; graphs are stratified by $\pi^{\bar{e}}$, $SE_{A\cap B}$, and the RR.

First, under the null hypothesis of a SE ratio equal to 1, the proposed test has a $Pr_{rej}$ close to the significance level 0.05. Then, the power of the test depends on the factors of the simulation design in the expected way:

- positive association with the size of the validation sample (strong)
- positive association with the distance of the SE ratio $SE^{e}/SE^{\bar{e}}$ from the null hypothesis (strong)
- positive association with the baseline SE in the unexposed (strong)
- negative association with the SE of the intersection $SE_{A\cap B}$ (strong)
- positive association with the distance of the RR from 1 (moderate)
- negative association with the prevalence in the unexposed $\pi^{\bar{e}}$ (weak)

Now we explore the cases when the power is large, namely, exceeding 0.8. For a validation sample of size 750, the power is always large if the SE ratio is 0.6 or 1.67, whereas, for values closer to 1, the power is large only for some configurations with $SE_{A\cap B} = 0$. For a validation sample of size 500, the performance worsens, though the power is still large for most configurations if the SE ratio is 0.6 or 1.67. Finally, for a validation sample of size 250, the performance further worsens, so the power is large only for some configurations if the SE ratio is 0.6 or 1.4, whereas it is quite low if the SE ratio is closer to 1.

## Adjusting the misclassification bias for number of cases, risk, risk ratio, and risk difference

In this section, we discuss a situation in which we have an observed outcome Y and have conducted validation studies for an indicator A with high SP and a screening algorithm B, obtaining estimates of the PPV for both indicators. The issue is how this information can be used to adjust measures of the number of cases and, as long as the PPVs have been estimated across exposure strata, to adjust the risk ratio and the risk difference.

In this section, we use the term "risk" to indicate a measure of risk where the numerator is the number of cases, and the calculation of the denominator does not depend on the detection of cases. This is the case for the prevalence, the cumulative incidence, and (with approximation) the rate of a rare event: in the latter case, indeed, the risk is the number of new cases divided by the person-time still at risk, which depends on the detection of cases. However if the event is rare, a small variation in the detection of new cases has a negligible impact on the person time at risk.

### Number of cases and risk

The number of cases observed in the population $N_A$ differs from the true number $N_Y$. If we know $PPV_A$ and $SE_A$, a straightforward correction is possible:

$$N_Y = N_A \times \frac{PPV_A}{SE_A} \tag{8}$$

This is because

$$\begin{aligned} N_Y &= TP + FN \\ &= (TP + FN) \times \frac{(TP+FP)}{TP} \times \frac{TP}{(TP+FP)} \\ &= (TP + FP) \times \frac{(TP+FN)}{TP} \times \frac{TP}{(TP+FP)} \\ &= N_A \times \frac{1}{SE_A} \times PPV_A \end{aligned}$$

Denoting with $R_Y$ a measure of the risk for Y in the population and with $R_A$ its measure through A, and dividing each member of eqn (8) by the common denominator, we obtain the following formula:

$$R_Y = R_A \times \frac{PPV_A}{SE_A} \tag{9}$$

When we can assume that $SE_{A\cup B} = 1$, the number of cases and the true risk are then obtained as follows:

$$N_Y = N_A \times \frac{PPV_A}{1 - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A + P_B \times PPB_B - P_{A\cap B} \times PPV_{A\cap B}}} \tag{10}$$

and

$$R_Y = R_A \times \frac{PPV_A}{1 - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A + P_B \times PPB_B - P_{A\cap B} \times PPV_{A\cap B}}} \tag{11}$$

In cases where it cannot be assumed that $SE_{A\cup B} = 1$, a lower bound for the number of cases and for the risk is derived:

$$N_Y \geq N_A \times \frac{PPV_A}{1 - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A + P_B \times PPB_B - P_{A\cap B} \times PPV_{A\cap B}}} \tag{12}$$

and

$$R_Y \geq R_A \times \frac{PPV_A}{1 - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A + P_B \times PPB_B - P_{A\cap B} \times PPV_{A\cap B}}} \tag{13}$$

For example, in the Entresto study, we can adjust $N_Y$ as follows. Because $A \cap B = \varnothing$, the correction factor in is

$$\frac{\hat{P}PV_A}{1 - \frac{N_B \times \hat{P}PV_B}{N_A \times \hat{P}PV_A + N_B \times \hat{P}PV_B}} = 1.45 \tag{14}$$

Thus, the true number of cases is at least 1.45 times higher than $N_A = 34$:

$$N_Y \geq N_A \times 1.45 = 49.3 \tag{15}$$

Similarly, an estimate of risk for the study based on A would underestimate the true risk by at least 45%.

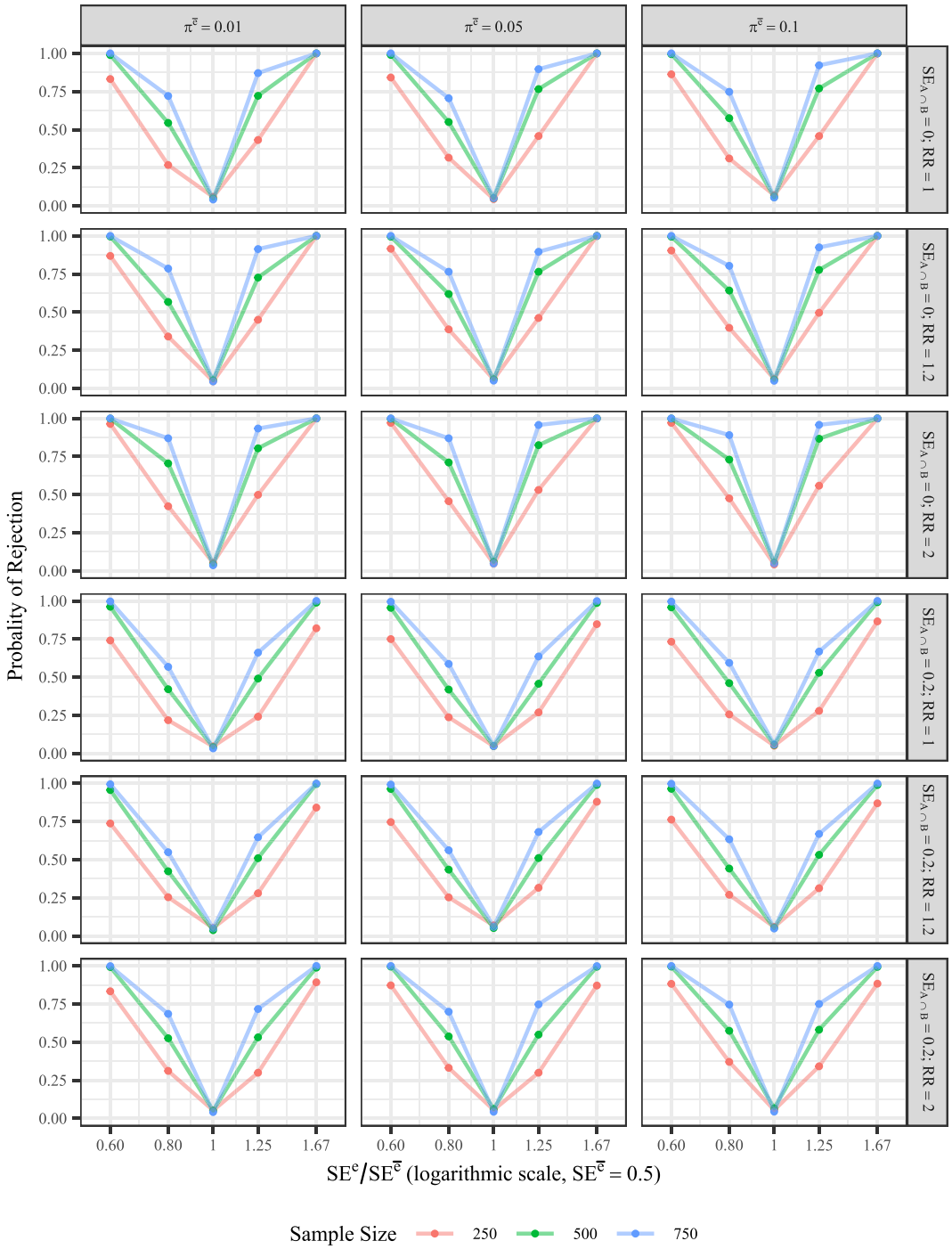


**Figure 1.** Proportion of times when the null hypothesis is rejected ($Pr_{rej}$). Different scenarios ($n = 18$) are presented in which sensitivity (SE) $SE_{A\cap B}$, relative risk (RR), and $\pi^{\bar{e}}$ vary. Each panel presents the power of the test power across 3 different sample sizes: 250, 500, and 750. $SE^{\bar{e}}$ is set to 0.5.

### *Risk ratio*

The true risk ratio is based on the outcome Y:

$$RR_Y = \frac{N_Y^e/N^e}{N_Y^{\bar{e}}/N^{\bar{e}}} \tag{16}$$

In database studies, it is approximated using the observed indicator A:

$$RR_A = \frac{N_A^e/N^e}{N_A^{\bar{e}}/N^{\bar{e}}} \tag{17}$$

Brenner and Gefeller[7] proved the following relationship:

$$RR_Y = RR_A \frac{PPV_A^e}{PPV_A^{\bar{e}}} \frac{SE_A^{\bar{e}}}{SE_A^e} \tag{18}$$

Note this formula also holds for rate ratios as long as the event is rare.

The common approach to adjust for measurement error is to assume that the SE is nondifferential across exposure groups

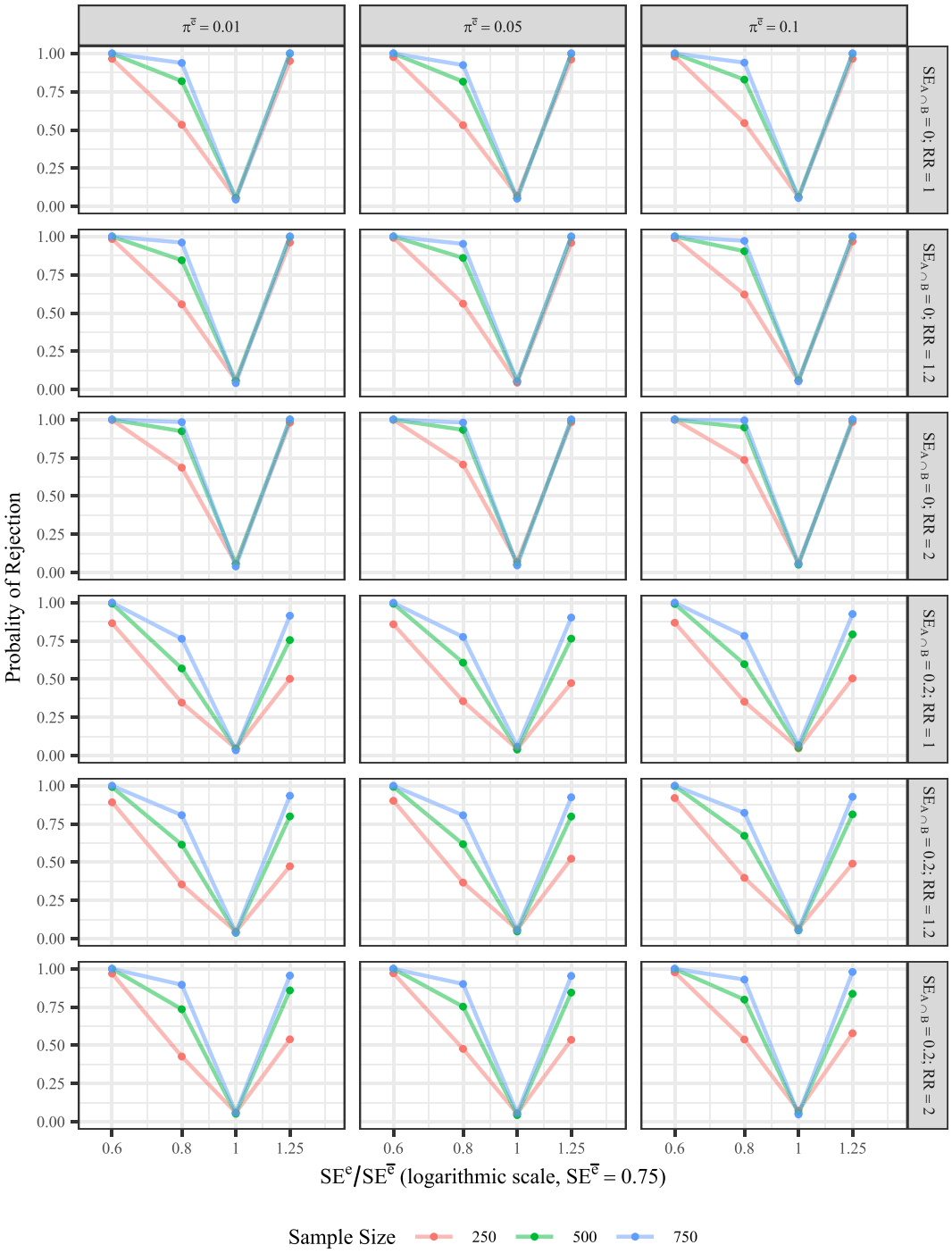


**Figure 2.** Proportion of times when the null hypothesis is rejected ($Pr_{rej}$). Different scenarios ($n = 18$) are presented in which sensitivity (SE) $SE_{A\cap B}$, relative risk (RR), and $\pi^{\overline{e}}$ vary. Each panel presents the power of the test power across 3 different sample sizes: 250, 500, and 750. $SE^{\overline{e}}$ is set to 0.75.

($SE^e_A = SE^{\overline{e}}_A$) so that the last factor in eqn (18) disappears and the correction only needs the PPVs of $A$. If the screening algorithm $B$ is available and it can be assumed that $SE_{A\cup B}$ is nondifferential ($SE^e_{A\cup B} = SE^{\overline{e}}_{A\cup B}$), the statistical test introduced in the previous section can be used to test the hypothesis $SE^e_A = SE^{\overline{e}}_A$ of nondifferential SE and make the results more robust. If the test fails, but it still can be assumed that $SE_{A\cup B}$ is nondifferential, then the correction (18) becomes

$$RR_Y = RR_A \times \frac{P^e_A PPV^e_A + P^e_B PPV^e_B - P^e_{A\cap B} PPV^e_{A\cap B}}{P^{\overline{e}}_A PPV^{\overline{e}}_A + P^{\overline{e}}_B PPV^{\overline{e}}_B - P^{\overline{e}}_{A\cap B} PPV^{\overline{e}}_{A\cap B}} \quad (19)$$

### *Risk difference*

The risk difference is defined as

$$RD_Y = \frac{N^e_Y}{N^e} - \frac{N^{\overline{e}}_Y}{N^{\overline{e}}}$$

By eqn (8), the risk difference can be calculated through the validity indexes of an indicator $A$ as follows:

$$RD_Y = \frac{PPV^e_A}{SE^e_A} \times \frac{N^e_A}{N^e} - \frac{PPV^{\overline{e}}_A}{SE^{\overline{e}}_A} \times \frac{N^{\overline{e}}_A}{N^{\overline{e}}} \quad (20)$$

If A is chosen to have a very large PPV, this may come at the expense of the SE. shows that this may induce a significant underestimation of the risk difference: the PPV of the screening indicator B can help quantify this.

Multiple scenarios are possible depending on the assumptions that can be made on A and on the PPVs that are available or can be estimated. We first describe a simple scenario (scenario 1) and prove a lower bound for the underestimation of the risk difference based on $PPV_B$. We then discuss a set of formulas that can be used when this simple scenario is not valid (scenarios 2-5) and prove them in Appendix S4.

All scenarios, except the second, assume the existence of the screening indicator B. In scenarios 1 and 3, the condition that $SE_A$ is nondifferential is required. This assumption can be justified on theoretical grounds or, if the $PPV_B$ can be estimated across exposure strata, it can be verified through the nondifferentiality test of Section "A test for non-differential sensitivity".

**Scenario 1:** $SE_A$ nondifferential ($SE_A^e = SE_A^{\bar{e}}$) and $PPV_A^e = PPV_A^{\bar{e}} = PPV_A = \mathbf{1}$, and $P_{A\cap B} = \mathbf{0}$. This is the case when indicator A has been chosen to be highly specific. The statement that $PPV_A = 1$ in both exposure strata may be an assumption or may be supported by a validation study. Replacing the assumptions in proves the following formula:

$$RD_Y = \frac{1}{SE_A} \times RD_A$$

where $RD_A$ is the observed risk difference

$$RD_A = \frac{N_A^e}{N^e} - \frac{N_A^{\bar{e}}}{N^{\bar{e}}}$$

Hence, using the upper bound for $SE_A$ give in eqn (3), we get the following lower bound for the estimator of the risk difference:

$$\mathbf{sign}\,(RD_Y) = \mathbf{sign}\,(RD_A) \tag{21}$$

$$|\,RD_Y\,| \geq \left|\,\frac{P_A + P_B \times PPV_B}{P_A} RD_A\,\right| = \left|\,\left(1 + \frac{P_B \times PPV_B}{P_A}\right) RD_A\,\right| \tag{22}$$

Note that this formula does not require that $PPV_B$ is estimated across exposure strata.

**Scenario 2:** $SE_A$ nondifferential ($SE_A^e = SE_A^{\bar{e}} = SE_A$). This is a scenario in which the auxiliary indicator B is not available but estimation stratified by exposure of the PPV of A can be obtained. Then, the following lower bound can be used:

$$\mathbf{sign}(RD) = \mathbf{sign}\left(PPV_A^e \times P_A^e - PPV_A^{\bar{e}} \times P_A^{\bar{e}}\right) \tag{23}$$

$$|\,RD\,| \geq \left|\left(PPV_A^e \times P_A^e - PPV_A^{\bar{e}} \times P_A^{\bar{e}}\right)\right| \tag{24}$$

**Scenario 3:** $SE_A$ nondifferential ($SE_A^e = SE_A^{\bar{e}}$) and $PPV_A^e = PPV_A^{\bar{e}} = PPV_A = \mathbf{1}$. This is the generalization of scenario 1 to the case when $P_{A\cap B} > \mathbf{0}$. Again, a lower bound can be provided where the PPV of B can be estimated in the general population (and, therefore, be retrieved from a previous study, if available):

$$\begin{array}{ll} \mathbf{sign}(RD) = & \mathbf{sign}\left(P_A^e - P_A^{\bar{e}}\right) \\ |\,RD\,| \geq & \left|\,\frac{P_A + P_B PPV_B - P_{A\cap B}}{P_A - 2P_{A\cap B}}\left(P_A^e - P_A^{\bar{e}}\right)\right| \end{array} \tag{25}$$

If $A \cap B = \varnothing$,then:

$$|\,RD\,| \geq \left|\,\frac{P_A + P_B PPV_B}{P_A}\left(P_A^e - P_A^{\bar{e}}\right)\right| \tag{26}$$

**Scenario 4:** $SE_{A\cup B}$ nondifferential ($SE_{A\cup B}^e = SE_{A\cup B}^{\bar{e}} = SE_{A\cup B}$). This is the generalization of scenario 1 to the case when the assumption that there are no false positives in A cannot be made, and $P_{A\cap B} \geq 0$. Then, the PPV of A, B, and (if $P_{A\cap B} > \mathbf{0}$) $A \cap B$ must be estimated across exposure strata, and the following lower bound holds:

$$\begin{aligned} \mathbf{sign}(RD) = \mathbf{sign}\,\big(&\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right) \\ &- \left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right)\big) \end{aligned} \tag{27}$$

$$\begin{aligned} |\,RD\,| \geq |\,&\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right) \\ &- \left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right)\,| \end{aligned}$$

If, on top of this, the assumption $SE_{A\cup B} = \mathbf{1}$ can be made, we obtain the equation:

$$\begin{aligned} RD = \quad &\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right) \\ &- \left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right) \end{aligned} \tag{28}$$

**Scenario 5:** the PPV of A is greater than SE in the exposed group ($PPV_A^e > SE_A^e$). We finally give a formula that can be used when PPV of A and B are available only among nonexposed: when exposure is rare, such parameters can be assumed to be equal to those in the general population, that may be available from previous studies. The formula holds under the assumption that the PPV of A among exposed is higher than its SE, which is plausible if A can be built with convincingly high SP.

$$\mathbf{sign}(RD) = \mathbf{sign}\left(P_A^e - \frac{1}{SE_{A\cup B}}\left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right)\right) \tag{29}$$

$$|\,RD\,| \geq \left|\,P_A^e - \frac{1}{SE_{A\cup B}}\left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right)\right|$$

where $\mathbf{1}/SE_{A\cup B} \geq 1$.

## Discussion

When an event Y is the outcome of a database study, it is common practice to choose a very specific indicator A for Y, and either assume that it has no false positives or, if validation is possible, to estimate $PPV_A$ to account for false positives captured by A. In this article, we have elaborated on the notion of a screening indicator B aimed at capturing all (or nearly all) of the cases missed by A so that it has near-perfect SE while still being specific enough to exclude many noncases. We showed that, by including the estimation of $PPV_B$ in the validation study, we can at least partially account for false negatives that A fails to identify. In Table 1, we summarize the opportunities provided by using a screening algorithm.

Recently, the estimation of the SE of indicators was recommended by regulatory authorities as a means to improve the quality of evidence generated for regulatory purposes.[6] Although this recommendation is considered very arduous to comply with,[8] we introduced a method that partially meets the recommendation.

When estimating the occurrence of Y, we provided a formula to adjust, or provide a lower bound for, the number of cases,

**Table 1.** Scenarios for estimating sensitivity of A, cases, risk, risk ratio, risk difference, and testing for nondifferentiality.

| | **No screening algorithm** | **With screening algorithm** | | |
|---|---|---|---|---|
| **Population not stratified per exposure** | **$PPV_A$ available** | **$PPV_A$, $PPV_B$, $PPV_{A\cap B}$ available** | | |
| | – | $SE_{A\cup B}=1$ | $SE_{A\cup B}<1$ | |
| Sensitivity of A, no. of cases risk | Not estimable | Unbiased | Upper bound | |
| | Underestimated | Unbiased | Reduced underestimation | |
| | Underestimated | Unbiased | Reduced underestimation | |
| **Population stratified per exposure** | **$PPV_A$ available in both strata** | **$PPV_A$, $PPV_B$, $PPV_{A\cap B}$ available in both strata** | | |
| | – | $SE_{A\cup B}=1$[a] | $SE^e_{A\cup B}=SE^{\overline{e}}_{A\cup B}<1$ | $SE^e_{A\cup B}\neq =SE^{\overline{e}}_{A\cup B}$ |
| Risk ratio | Unbiased if $SE^{\overline{e}}_A=SE^e_A$ | Unbiased | Unbiased | Unbiased if $SE^{\overline{e}}_A=SE^e_A$ |
| Risk difference | Biased | Unbiased | Lower bound in absolute value[b] | Biased |
| Differentiality of A | Not testable | Testable | Testable | Not testable |

Abbreviations: PPV, positive predictive value; SE, sensitivity.
[a]In this scenario, sensitivity across exposure groups can be estimated, and measures of association can be adjusted using bias analysis.
[b]This corresponds to scenario 4. Other scenarios are reported and allow for different improvements.

the prevalence, the cumulative incidence, and the incidence rate of Y (the latter being valid only for rare events). For association studies, we introduced several tools to address differential misclassification.

Choosing an appropriate screening algorithm may depend on several factors. An extreme example is when $B=\overline{A}$, meaning B retrieves all units not identified by A (ie, $A\cup B$ covers the whole population; in this case, $SE_{A\cup B}=1$). However, estimating the PPV of B in this scenario is not convenient, because the expense is similar to that of directly estimating the SE of A. Conversely, an overly specific algorithm may fail to detect any individuals not already identified by the main algorithm (no false negatives of A in B). Here, the screening algorithm provides no additional information to the main algorithm and is not useful for SE estimation.

A key factor to be taken into account is the sample size of the validation study which is, for feasibility, typically on the order of hundreds (this is why we chose this range in our simulations). To be useful, the screening indicator must identify at least a few true cases outside of A within the validation sample. To this end, $SP_B$ should be large enough depending on the prevalence (this is why we set $SP_B=1-\pi$ in our simulations). The choice of B, which has to be done before the validation study, should be based on the clinical plausibility that cases are found outside of A.

In our simulation studies, we explored different scenarios concerning the validation indices of B, even though we consistently remained within scenarios where prevalence and $SP_B$ are linked by the relationship $SP_B=1-\pi$. The power of the test does not primarily depend on the specifics of B (as long as some true positives are identified), nor does the estimation of $SE_A$. The power of the test mainly depends on the SE ratio across exposure groups, the intersection between A and B, and the sample size. Regarding the estimation of SE, the uncertainty of the estimate depends on the uncertainty in estimating the PPVs, which, in turn, primarily depends on the sample size.

A set of important examples on how the screening indicator B can be defined was constructed in the Vaccine COVID-19 Monitoring Readiness (ACCESS) study[1] during the recent pandemic emergency. Ahead of the vaccination campaign, the ACCESS study was tasked by the European Medicines Agency to calculate background rates for 41 potential adverse events of special interest. Using the medical definitions of each event, the tool Codemapper,[26] which leverages the Unified Medical Language System, provided a comprehensive list of potentially relevant diagnostic codes. These codes were categorized by medical experts as "narrow" if they strictly imply the medical condition of interest, or as "possible" if they could also denote other conditions with similar clinical presentation. Such lists are stored in the Zenodo public repository; for example, see the definition of meningoencephalitis.[27] The list of codes tagged as narrow was used in ACCESS as algorithm A to identify background rates of the event of special interest. The possible codes list was not used. In future applications, it could serve as the screening algorithm B. Even if $SE_{A\cup B}=1$ is not guaranteed (eg, patients with nonsevere adverse events of special interest may not access the health care system), validating a sample of cases detected by possible codes and using our formulas would still reduce the underestimation of the incidence when compared with the incidence estimated using only narrow codes. This would increase the number of expected cases in vaccinated cohorts, thereby reducing the risk of triggering false alerts of an adverse reaction to vaccines in an observed-to-expected analysis.[2,28]

Our tools allow the application of the following strategy. When estimating the fitness-for-purpose of a data source to assess the occurrence of an event Y, a specific indicator A for Y can be sought alongside a screening indicator B. The SE of a A in the population can be estimated by validating both a sample of A and a sample of B, using the methodology described in Section "Sensitivity of indicators". This information can be used immediately to assess the number of cases of Y in the population and in subpopulations. Moreover, it can be used in a later stage when association studies are conducted with Y as an outcome.

If the risk ratio is the measure of association, knowledge of the SE of A in the population can be used to evaluate the impact of differential misclassification: if A has low SE in the population, the impact of differential SE can be high, because SE may become substantially higher in special populations, such as those exposed to a new medication.

Differential SE is not a trivial matter. In safety studies, it may occur when studying suspected adverse reactions to a medicinal product because a clinician assessing the symptoms of a patient may use their exposure to the medicine in the diagnostic process, and the knowledge that the outcome is a possible adverse reaction may lead to record a more specific diagnostic code. Our tools are based on estimates of the PPVs of A and B across exposure strata. We introduced a statistical test for nondifferential SE based on the assumption that $SE_{A\cup B}$ is nondifferential and exposure is not misclassified.

Our test showed good power in detecting modest departures from nondifferentiality with a validation sample of size 500,

retrieved from cases detected by $A$, $B$, and $A \cap B$ (if nonempty), independently on the prevalence of $Y$. This test can be used to decide whether association measures between $Y$ and exposure are unbiased if adjusted by the PPV of $A$. If the test fails, we introduced a formula to adjust RRs and rate ratios for a rare $Y$, based on the same PPVs that were used for the test.

When applying our formulas to the case of association studies, we release the assumption that sensitiviy of $A$ is nondifferential and replace it with the request that $A \cup B$ is nondifferential. This is automatically satisfied when $B$ captures all true cases missed by $A$ but can be assumed in other situations. Indeed, note that when the physician assesses a set of symptoms, use of medications, especially in case of a suspect adverse reaction, can be used in the process of diagnosis to obtain a more specific diagnosis. Those without the medication presenting the same symptoms would then be registered with a less specific diagnosis. This is a case when the SE of the specific indicator $A$ would be differential, whereas the SE of $A \cup B$ would not. A practical example would be the case of angioedema and angiotensin-converting enzyme inhibitors: because angioedema is a known adverse reaction to angiotensin-converting enzyme inhibitors, users (ie, individuals in the exposed stratum) will be more likely to be recorded with the specific code ($A$), rather than being recorded as hypersensitive ($B$). Therefore, $SE^e_A$ is expected to be higher than $SE^{\bar{e}}_A$. However, symptoms are the same, so the overall likelihood of the patient to access the emergency room is the same across exposure strata, and patients missed by $A$ are picked by $B$; therefore, the union of the 2 indicators is expected to have nondifferential SE.

The knowledge of $P_A$, $P_B$, and $P_{A\cap B}$ across exposure strata can be used to run a simulation of the power of our test, as done in Section "A test for non-differential sensitivity", and to decide the sampling strategy for a validation study stratified by exposure. If the test fails, the results from the validation study can also be used to adjust the estimates. In addition, a test for the equality of the SEs is useful even if the SEs can be estimated in the exposure groups separately, because this entails a loss of precision as compared with pooled estimation, in analogy to the case of comparing 2 means using a $t$ test with unequal variances when the population variances are equal.[29]

If the association measure of interest is the risk difference, then knowledge of the SE of $A$ and the context of $Y$ can be used to assess whether 1 of the scenarios of Section "Risk difference" applies to adjust for misclassification. If required by the scenario, a validation study stratified per exposure may also be conducted in this case.

A limitation of our test is that it does not account for errors in the exposure. Specifically, we did not consider the potential dependency between outcome misclassification and exposure misclassification. In observational studies, systematic differences in determining exposures may also affect the accuracy of diagnoses, even if only marginally. Research has shown that this scenario can lead to substantial bias in the results, even with minor levels of misclassification. In a future development, the test can be adapted to address also exposure misclassification.

In addition, we did not provide estimates of the uncertainty around our formulas. However, the uncertainty can always be assessed through simulation-based approaches such as the bootstrap. Another limitation is that, in the simulations, we did not explore a range of options for the SP of $A$ and $B$. Such scenarios need further research.

Areas for further research could include extending the proposed method to nonrare events and hazard ratios, as well as incorporating the method in statistical modeling possibly with weighted estimation.

## Conclusions

Whenever an auxiliary screening indicator can be defined to complement a primary indicator for an event, estimates of the PPV of both indicators provide tools to reduce the bias in estimating the number of cases, prevalence, cumulative incidence, rate (if the event is rare), and, in the case of association studies, risk ratio and risk difference. They also allow to test for nondifferential SE.

Although direct estimation of the SE is often infeasible, this novel methodology improves evidence based on data obtained from reuse of existing databases, which may prove critical for regulatory and public health decisions. Our work encourages in future studies a more widespread use of validation studies to support or reject assumptions on misclassification.

## Funding

The ConcePTION project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no. 821520. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations.

## Conflict of interest

None declared.

## Disclaimer

The research leading to the results reporter here was conducted as part of the ConcePTION consortium. This article only reflects the personal views of the stated authors.

## References

1. Willame C, Dodd C, Durán CE, et al. Background rates of 41 adverse events of special interest for COVID-19 vaccines in 10 European healthcare databases-an access cohort study. *Vaccine*. 2023;41(1):251-262. https://doi.org/10.1016/j.vaccine.2022.11.031
2. Durand J, Dogné JM, Cohet C, et al. Safety monitoring of covid-19 vaccines: perspective from the European Medicines Agency. *Clin Pharmacol Ther*. 2023;113(6):1223-1234. https://doi.org/10.1002/cpt.2828
3. Thurin NH, Pajouheshnia R, Roberto G, et al. From inception to conception: genesis of a network to support better monitoring and communication of medication safety during pregnancy and breastfeeding. *Clin Pharmacol Ther*. 2022;111(1):321-331. https://doi.org/10.1002/cpt.2476
4. Labgold K, Stanhope KK, Joseph NT, et al. Validation of hypertensive disorders during pregnancy: ICD-10 codes in a high-burden southeastern United States hospital. *Epidemiology*. 2021;32(4):591-597. https://doi.org/10.1097/EDE.0000000000001343
5. Bollaerts K, Rekkas A, de Smedt T, et al. Disease misclassification in electronic healthcare database studies: deriving validity indices—a contribution from the advance project. *PLoS One*. 2020;15(4):e0231333. https://doi.org/10.1371/journal.pone.0231333
6. Food and Drug Administration. *Real-world data: assessing electronic health records and medical claims data to support regulatory*

*decision-making for drug and biological products draft guidance for industry*. Food and Drug Administration; 2021. Accessed October 18, 2021. Available at: https://www.fda.gov/media/152503/download
7. Brenner H, Gefeller O. Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. *Am J Epidemiol*. 1993;138(11):1007-1015. https://doi.org/10.1093/oxfordjournals.aje.a116805
8. Girman CJ, Ritchey ME, Re VL III. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. *Pharmacoepidemiol Drug Saf*. 2022;31(7):717-720. https://doi.org/10.1002/pds.5444
9. Hall GC, Lanes S, Bollaerts K, et al. Outcome misclassification: impact, usual practice in pharmacoepidemiology database studies and an online aid to correct biased estimates of risk ratio or cumulative incidence. *Pharmacoepidemiol Drug Saf*. 2020;29(11):1450-1455. https://doi.org/10.1002/pds.5109
10. Chung CP, Rohan P, Krishnaswami S, et al. A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. *Vaccine*. 2013;31(Suppl 10):K41-K61. https://doi.org/10.1016/j.vaccine.2013.03.075
11. Schneider G, Kachroo S, Jones N, et al. A systematic review of validated methods for identifying erythema multiforme major/minor/not otherwise specified, Stevens-Johnson syndrome, or toxic epidermal necrolysis using administrative and claims data. *Pharmacoepidemiol Drug Saf*. 2012;21(S1):236-239. https://doi.org/10.1002/pds.2331
12. Schneider G, Kachroo S, Jones N, et al. A systematic review of validated methods for identifying hypersensitivity reactions other than anaphylaxis (fever, rash, and lymphadenopathy), using administrative and claims data. *Pharmacoepidemiol Drug Saf*. 2012;21(S1):248-255. https://doi.org/10.1002/pds.2333
13. Tamariz L, Harkins T, Nair V. A systematic review of validated methods for identifying venous thromboembolism using administrative and claims data. *Pharmacoepidemiol Drug Saf*. 2012;21(S1):154-162. https://doi.org/10.1002/pds.2341
14. Tamariz L, Harkins T, Nair V. A systematic review of validated methods for identifying ventricular arrhythmias using administrative and claims data. *Pharmacoepidemiol Drug Saf*. 2012;21(S1):148-153. https://doi.org/10.1002/pds.2340
15. Re VL 3rd, Carbonari DM, Jacob J, et al. Validity of ICD-10-CM diagnoses to identify hospitalizations for serious infections among patients treated with biologic therapies. *Pharmacoepidemiol Drug Saf*. 2021;30(7):899-909. https://doi.org/10.1002/pds.5253
16. Jonsson Funk M, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep*. 2014;1(4):175-185. https://doi.org/10.1007/s40471-014-0027-z
17. Yland JJ, Wesselink AK, Lash TL, et al. Misconceptions about the direction of bias from nondifferential misclassification. *Am J Epidemiol*. 2022;191(8):1485-1495. https://doi.org/10.1093/aje/kwac035
18. Jurek AM, Greenland S, Maldonado G, et al. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol*. 2005;34(3):680-687. https://doi.org/10.1093/ije/dyi060
19. Jurek AM, Greenland S, Maldonado G. Brief report: how far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null? *Int J Epidemiol*. 2008;37(2):382-385. https://doi.org/10.1093/ije/dym291
20. Lanes S, Beachler DC. Validation to correct for outcome misclassification bias. *Pharmacoepidemiol Drug Saf*. 2023;32(6):700-703. https://doi.org/10.1002/pds.5601
21. Hempenius M, Groenwold RH, de Boer A, et al. Drug exposure misclassification in pharmacoepidemiology: sources and relative impact. *Pharmacoepidemiol Drug Saf*. 2021;30(12):1703-1715. https://doi.org/10.1002/pds.5346
22. Gini R, Dodd CN, Bollaerts K, et al. Quantifying outcome misclassification in multi-database studies: the case study of pertussis in the advance project. *Vaccine*. 2020;38(Suppl 2):B56-B64. https://doi.org/10.1016/j.vaccine.2019.07.045
23. Limoncella G, Girardi A, Hyeraci G, et al. Validation to estimate sensitivity along with positive predictive value: A case study. *Pharmacoepidemiol Drug Saf*. 2022;31(S2):440.
24. European Medicines Agency. Administrative details [Internet]. EMA; [cited June 29, 2025]. Available from: https://catalogues.ema.europa.eu/node/3321/administrative-detailsEUPAS18214
25. Efron B. Bootstrap methods: another look at the jackknife. In: Kotz S, Johnson NL, eds. *Breakthroughs in Statistics: Methodology and Distribution*. Springer; 1992:569-593.
26. Becker BF, Avillach P, Romio S, et al. Codemapper: semi-automatic coding of case definitions. A contribution from the advance project. *Pharmacoepidemiol Drug Saf*. 2017;26(8):998-1005. https://doi.org/10.1002/pds.4245
27. van Wijngaarden P, Belbachir L, Durán C, et al. *ACCESS-Background rate of adverse events- definition-(Meningo)encephalitis*. Zenodo; 2021 https://doi.org/10.5281/zenodo.5236137.
28. Gordillo-Marañón M, Candore G, Hedenmalm K, et al. Lessons learned on observed-to-expected analysis using spontaneous reports during mass vaccination. *Drug Saf*. 2024;47(7):607-615. https://doi.org/10.1007/s40264-024-01422-8
29. Moser BK, Stevens GR, Watts CL. The two-sample t test versus Satterthwaite's approximate F test. *Commun Stat Theory Methods*. 1989;18(11):3963-3975. https://doi.org/10.1080/03610928908830135




*American Journal of Epidemiology*, 2025, **194**, 2570–2579
https://doi.org/10.1093/aje/kwae423
Original Contribution