

University of Veterinary Medicine, Budapest
Doctor School of Veterinary Science

Handling misclassification in statistical analysis

PhD Dissertation

Dr. Hársfalvi Péter

2025

University of Veterinary Medicine, Budapest
Doctor School of Veterinary Science

Handling misclassification in statistical analysis

PhD Dissertation

Dr. Hársfalvi Péter

2025

Supervisor:

.....

Dr. Reiczigel Jenő

Associate Professor

Department of Biostatistics

University of Veterinary Medicine Budapest, Hungary

Copy ... of four.

.....

Dr. Hársfalvi Péter

Table of contents

Table of contents	3
1 Profile likelihood confidence interval for the prevalence assessed by an imperfect diagnostic test	15
1.1 BACKGROUND	15
1.2 METHOD	16
1.3 RESULTS AND DISCUSSION	19
1.3.1 <i>Comparison of PLCI to the Lang-Reiczigel CI</i>	19
1.3.2 <i>Comparison of PLCI to the Flor interval</i>	22
1.3.3 <i>Application examples</i>	22
2 Confidence limits for risk differences and risk ratios adjusted for estimated sensitivity and specificity	24
2.1 BACKGROUND	24
2.2 METHOD	25
2.3 DISCUSSION	28
2.3.1 <i>Evaluation of the method performance</i>	28
2.3.2 <i>Application examples</i>	34
3 The effect of misclassification on sample size for one and two-sample tests with binary endpoints	37
3.1 BACKGROUND	37
3.2 METHODS	40
3.2.1 <i>Introduction of the test methods</i>	40
3.2.2 <i>Study Settings</i>	43
3.3 RESULTS AND DISCUSSION	45
3.3.1 <i>Study Results</i>	45
3.3.1 <i>Case Studies</i>	49
4 Applying the promising zone method for sample size re-estimation in clinical trials when the binomial endpoint is based on a diagnostic test	51
4.1 BACKGROUND	51
4.2 METHODS	52
4.3 RESULTS AND DISCUSSION	54
4.3.1 <i>Example - a medical device trial for spinal disorder</i>	54
4.3.2 <i>General thoughts on the impact of misclassification</i>	56
5 Logistic regression with covariate-dependent probability of misclassification	59
5.1 BACKGROUND	59
5.2 METHODS	61
5.2.1 <i>Description of the model</i>	61
5.2.2 <i>Parameter estimation</i>	62
5.2.3 <i>Sub-model testing</i>	63
5.2.4 <i>Identifiability issues</i>	65
5.3 RESULTS AND DISCUSSION	68
5.3.1 <i>Power assessment</i>	68
5.3.2 <i>Applications</i>	72
5.3.3 <i>Discussion</i>	78
References	83

Notations

List of acronyms

AIC	Akaike's Information Criterion
CI	confidence interval
CP	coverage probability
CPo	conditional power
EL	expected length
ELISA	enzyme-linked immunosorbent assay
EM	expectation-maximization
HDI	highest density interval
LCL	lower confidence limit
LMES	level of the management and environmental sanitation
LRT	likelihood ratio test
PLCI	profile likelihood confidence interval
RD	risk difference
RR	risk ratio
Se	Sensitivity
Sp	Specificity
TB	tuberculosis
UCL	upper confidence limit

List of symbols

P	probability
θ	the parameter of interest, may be vector-valued
θ_0	the parameter of interest under the null hypothesis
ν	some nuisance parameters, may be vector-valued
H_0	null hypothesis
H_1	alternative hypothesis
x	observed sample
Q	the ratio of likelihoods
L	likelihood
L_0	the likelihood maximized over the nuisance parameter ν while keeping θ at its null value θ_0
L_s	likelihood maximized over both θ and ν
α	type I error
$\chi_{1,1-\alpha}^2$	the $1 - \alpha$ quantile of the chi-square distribution having 1 degree of freedom
p	prevalence of a disease with p_1 and p_2 being the prevalences in groups 1 and 2 in a two-group study
p_0	prevalence of a disease under the null hypothesis
p_a	prevalence of a disease under the alternative hypothesis
Se	sensitivity
Sp	specificity
n_{se}	size of the sample used for the estimation of sensitivity
n_{sp}	size of the sample used for the estimation of specificity
n/N	size of the sample
k_{se}	number of correctly diagnosed out of n_{se}
k_{sp}	number of correctly diagnosed out of n_{sp}
k	number of correctly diagnosed out of n

z_{crit}^2	percentile of the standard normal distribution corresponding to a certain (two-sided) confidence level
\hat{t}'_i	cell frequency, response rate of observed risk
\hat{R}_i	estimated risk in group i
l_i	lower confidence limit for the estimated parameter in group i
u_i	upper confidence limit for the estimated parameter in group i
X	random variable
C	critical region
δ	treatment difference
ϕ	cumulative distribution function of the standard normal distribution
Y_{true}	true response in the logistic model
Y_{obs}	observed response in the logistic model
$\beta_{0...i}$	regression coefficients of the logistic regression model
$\gamma_{0...i}$	regression coefficients of the logistic regression model used for the modelling of the sensitivity

List of figures

- Figure 1** Coverage curves of 95% CIs for $n_{Se} = n_{Sp} = n = 100$. Blue curve: PLCI with adjustment, red curve: PLCI without adjustment, black curve: Lang-Reiczigel CI.
- Figure 2** Exact coverage rates for the 95% confidence interval of the RR with different sample sizes used for the estimation of sensitivity and specificity. a) $n_{Se} = 20$, $n_{Sp} = 20$. b) $n_{Se} = 50$, $n_{Sp} = 20$. c) $n_{Se} = 20$, $n_{Sp} = 50$ and d) $n_{Se} = 50$, $n_{Sp} = 50$. True Se and Sp are 0.9 in both groups. n_1 and n_2 are both selected to be 20.
- Figure 3** Exact coverage rates for the 95% confidence interval of the RD with different sample sizes used for the estimation of sensitivity and specificity. a) $n_{Se} = 20$, $n_{Sp} = 20$. b) $n_{Se} = 50$, $n_{Sp} = 20$. c) $n_{Se} = 20$, $n_{Sp} = 50$ and d) $n_{Se} = 50$, $n_{Sp} = 50$. True Se and Sp are 0.9 in both groups. n_1 and n_2 are both selected to be 20.
- Figure 4** Exact coverage rates for the 95% confidence interval of the RD with different true sensitivity and specificity values. a) $Se = Sp = 0.90$. b) $Se = Sp = 0.95$. c) $Se = Sp = 0.99$. $n_{Se} = 20$, $n_{Sp} = 20$ and n_1 and n_2 are both selected to be 20.
- Figure 5** Power of the exact binomial test as the function of the Se or Sp, while the other parameter is fixed at 1 with ($p_0 = 0.02$, $p_a = 0.05$, alternative = "two-sided", $n = 400$).
- Figure 6** Power of the exact binomial test is not a monotonic function of sample size ($p_0 = 0.5$, $p_a = 0.4$, alternative = "left-sided", $Se = Sp = 1$).
- Figure 7** Probability of each interim outcome zone with and without adjustment for two different true fusion rates (all results are based on 1.000.000 simulated trials).
- Figure 8** Adjusted and unadjusted Z1-values for different null hypotheses and different response rates, for $n_1=50$, $Se=0.9$ and $Sp=0.85$.
- Figure 9** Diagram of the recommended model comparisons.
- Figure 10** Estimated dependence of the prevalence of alcohol consumption on age, and dependence of Se (honest reporting of alcohol consumption) on how often the respondent admittedly rode a car with a drinking driver. The estimated Sp is 1, which means that those respondents who do not consume alcohol are assumed to report their status accurately.

Figure 11 Estimated age dependence of the prevalence of alcohol consumption assuming constant Se , that is, applying the Liu-Zhang model (left panel). The estimates of Se and Sp are $Se = 0.73$ and $Sp = 1$. Estimated age-dependence of the prevalence of alcohol consumption assuming no misclassification (right panel).

Figure 12 Results of our model and those of the Liu-Zhang model for the Czech Republic.

List of tables

- Table 1** Relative difference between the expected length of 95% two-sided PLCI and Lang-Reiczigel CI by true prevalence (values for true prevalence > 0.5 follow from symmetry). Negative values indicate that PLCI is shorter (length of the Lang-Reiczigel CI is regarded as 100%).
- Table 2** Length improvement of 95% two-sided PLCI compared to Lang-Reiczigel CI depending on sensitivity, specificity and true prevalence for $n = 1000$, $n_{Se} = n_{Sp} = 100$. Negative percentage values (with gray shading) indicate that PLCI is shorter (length of the Lang-Reiczigel CI is regarded as 100%). Darker shading indicates improvement greater than 5%.
- Table 3** Adjusted 95% CIs of RR for different n_{Se} and n_{Sp} values ($n_{Se} = n_{Sp}$), the observed response rates in the two arms are 30/60 and 40/80.
- Table 4** Adjusted 95% CIs of RR for different n_{Se} and n_{Sp} values ($n_{Se} \neq n_{Sp}$), the observed response rates in the two arms are 30/60 and 40/80.
- Table 5** Adjusted 95% CIs of RD for different n_{Se} and n_{Sp} values ($n_{Se} \neq n_{Sp}$), the observed response rates in the two arms are 30/60 and 40/80.
- Table 6** Null and assumed true probabilities for the one-sample test. Left- and right-tailed tests were evaluated separately.
- Table 7** The p_1 and p_2 pairs used in the evaluation of the two-sample test.
- Table 8** Sample sizes for the one-sample exact test with fixed Se and Sp for all alternatives in scenarios with equal sensitivity and specificity.
- Table 9** Sample size increase in % per group for the two-sample test with fixed Se and Sp in scenarios with equal Se and Sp for both groups For $Se_1=Se_2=Sp_1=Sp_2=1$ the required sample size is 110 as a baseline.
- Table 10** Sample sizes for the one-sample exact test with Se and Sp estimated from validation studies with different sample size, $p_0 = 0.01$ and $p_a = 0.1$ with $Se = 1$. The sample size needed for the scenarios with the same parameters but fixed Se and Sp are in the last row.

- Table 11** The range of differences between the adjusted and unadjusted CP (%) values for $n_2=200$, $Se=0.9$ and $Sp=0.85$.
- Table 12** Response rate ranges and ranges of linear predictor in italics used in the simulation experiments.
- Table 13** Power of detecting dependence of Y on X for various outcome probability and Se ranges and Sp in the 5-range experiments for a sample size of $N=1000$. Each number is estimated from 500 replications.
- Table 14** Age distribution of the respondents.
- Table 15** Each of the 15 items in Question 44 was tested on the Czech data, whether the misclassification probability $P(\text{reports always} \mid \text{not always votes})$ depends on the answer to it. The dependence was significant for 5 items, highlighted in bold. All these 5 items are related to „comme il faut” behavior, and 4 of them are associated with minor misconduct against the state.

Acknowledgement

I am deeply grateful to a multitude of people who have contributed to the successful completion of my dissertation.

First and foremost, I would like to extend my sincerest gratitude to my supervisor, Professor Jenő Reiczigel. Your insightful guidance, constructive criticism, and constant encouragement have been invaluable throughout the entire process. None of my academic work would have been possible without you.

I also wish to express my appreciation to the members of my dissertation committee: Dr. Júlia Singer, Dr. Tamás Ferenczi and Professor Jan Klaschka. Your feedback and suggestions have significantly improved the quality of our work.

Special thanks are due to Dr. Andrea Harnos, Dr. Zsolt Lang and Dr. Péter Fehérvári. Your encouragement and invaluable discussions have helped me a lot.

Finally, I am profoundly grateful to my family for their unwavering support.

Introduction

In medical research, behavioral sciences, and numerous other fields, accurate measurement is crucial for obtaining reliable results. However, measurements are often prone to errors, which can compromise the validity of research findings. When the measured variable is continuous, these errors are typically categorized as "measurement errors." In contrast, for categorical variables such as binary diagnostic statuses, the corresponding errors are referred to as "misclassification" (Grace, 2016; Gustafson, 2003). Misclassification arises when the recorded category does not accurately reflect the true status, often due to limitations in diagnostic tests or observer errors.

While the body of literature on handling measurement errors in continuous data is vast and well-developed, the same cannot be said for categorical data. Measurement errors in continuous, especially normally distributed, data have been extensively studied, with numerous techniques available for managing these issues (Carroll et al., 1995; Fuller, 2009; Klepper & Leamer, 1984). In contrast, the literature addressing the influence of misclassification and proposing methods to handle it during research planning or analysis is relatively sparse. Despite recent calls for more attention to misclassification, there remains a lack of comprehensive analytical methods tailored to specific or complex models.

The consequences of misclassification can be significant, leading to biased estimates, reduced statistical power, and erroneous conclusions. Given the widespread use of diagnostic tests in medical and epidemiological research, accounting for misclassification has become an increasingly critical concern. In particular, diagnostic tests often exhibit less than perfect sensitivity and specificity, resulting in the potential misclassification of disease statuses. As these diagnostic parameters are frequently estimated from small or moderate sample sizes, the uncertainty associated with them further exacerbates the problem.

A review of the literature reveals several approaches to addressing misclassification, but many are limited in scope or applicability. Existing methods often perform poorly when sample sizes are small or when sensitivity and specificity are high. Furthermore, some methods require complex computations, limiting their practical application. The literature still lacks robust techniques for handling misclassification in certain complex research scenarios, including logistic regression and adaptive clinical trial designs.

The overarching goal of this research is twofold. First, it aims to extend existing statistical methods that do not perform well in the presence of misclassification, making them suitable

for scenarios where misclassification may be present. Specifically, the research focuses on enhancing logistic regression models and developing new approaches for constructing confidence intervals.

Second, the research seeks to develop novel methods for addressing complex research problems involving misclassification. Key contributions include:

- A new profile likelihood confidence interval (PLCI) for estimating disease prevalence when sensitivity and specificity are known or estimated from independent validation samples.
- Sample size calculation methods for various tests involving binary data, accounting for the effects of misclassification.
- A sample size re-calculation method for the popular adaptive clinical trial design, the "promising zone" method.

By combining theoretical and simulation-based techniques, the research offers practical solutions that are computationally efficient and adaptable to real-world applications. The proposed methods aim to improve the accuracy and reliability of statistical inferences in the presence of misclassification, ultimately contributing to more robust research findings.

The thesis is divided into five distinct chapters, each addressing a specific statistical methodology:

1. **Profile Likelihood Confidence Interval for Disease Prevalence:** This chapter introduces the PLCI method and compares its performance with existing methods.
2. **Confidence Limits for Risk Differences and Risk Ratios:** The focus here is on constructing confidence intervals adjusted for estimated sensitivity and specificity.
3. **Effect of Misclassification on Sample Size:** This chapter explores the impact of misclassification on sample size requirements for one- and two-sample tests with binary endpoints.
4. **Promising Zone Method for Sample Size Re-estimation:** A novel approach for sample size re-calculation in adaptive clinical trials involving diagnostic tests.
5. **Logistic Regression with Covariate-Dependent Misclassification:** This chapter presents an advanced model for handling misclassification in logistic regression, addressing identifiability issues and parameter estimation.

Each chapter begins with an introduction to the background and motivation for the research, followed by a detailed presentation of the methods and a discussion of the findings. The

chapters conclude with subject-specific discussions that highlight the implications and potential applications of the research.

By addressing critical gaps in the literature and proposing innovative solutions, this dissertation contributes to the growing body of research on handling misclassification in statistical analysis. The methods developed here have the potential to improve the accuracy and reliability of findings in medical research, behavioral sciences, and other fields where categorical data play a central role. Through a combination of theoretical advancements and practical applications, this research aims to pave the way for more robust and reliable statistical methodologies.

1 Profile likelihood confidence interval for the prevalence assessed by an imperfect diagnostic test

The content of the chapter is based on my article with the same title, published in Preventive Veterinary Medicine in 2023.

1.1 Background

Estimating the prevalence of a disease is an essential task for most epidemiological studies. Instead of providing a single point estimate for the prevalence, it is more common to provide limits which are likely to include the parameter, called constructing an interval estimate for the prevalence of a disease. Multiple different methods for interval estimation are available but the most used conventional methods do not take potential misclassification into account. (Bland, 2015)

The device used in most epidemiological scenarios to distinguish between people in the population who have the disease and those who do not are called diagnostic tests. As most diagnostic tests used for the determination of disease status are prone to misclassification, it is important to incorporate this in the interval estimation of the prevalence as well.

When the outcome of an experiment is disease status and a diagnostic test is applied, the two usual measures of test quality are Sensitivity (Se), the proportion of correct diagnosis given the subject has the disease, and Specificity (Sp), the proportion of correct diagnosis given the subject does not have the disease (Yerushalmy, 1947):

$$Se = P(\text{test positive} \mid \text{diseased}) = 1 - P(\text{false negative}), \quad (1)$$

$$Sp = P(\text{test negative} \mid \text{non-diseased}) = 1 - P(\text{false positive}). \quad (2)$$

Many different solutions have been used for the adjustments of the disease prevalence point estimates and CIs obtained from the screening of the population. Rogan and Gladen proposed an adjustment for the point estimate of the disease prevalence (the so-called apparent prevalence) for the specific case when Se and Sp are considered as known values (Rogan & Gladen, 1978). They also introduced an asymptotic interval estimate for the true prevalence, which was based on normal approximation. Unfortunately, the interval had rather poor performance (Greiner & Gardner, 2000). Reiczigel et al. were the first to present a method to construct exact CIs for the same problem of Se and Sp considered as known parameters (Reiczigel et al., 2010).

Not long after, Lang and Reiczigel were the ones who also pointed out that considering Se and Sp estimates as fixed values (known parameters) might not be optimal, as this approach ignores the uncertainty in the estimates. When Se and / or Sp are estimated from moderate or small samples, it may lead to unreasonably optimistic CIs (Lang & Reiczigel, 2014). Instead, they proposed a method for taking uncertainty in Se and Sp into account by combining the previously mentioned Rogan and Gladen formula with an adjustment similar to the well-known “add 2 success and 2 failure” by Agresti and Coull (Agresti & Coull, 1998).

Flor et al. applied a Bayesian credible interval, the highest density interval (HDI) for the same problem. HDI is a range of parameter values with the highest posterior density containing a specified percentage of the posterior distribution’s probability mass. Thus, a 95% HDI is the shortest 95% credible interval with any parameter value outside the HDI considered less plausible than the values inside of it (Flor et al., 2020).

The advantage of the last two presented methods (Reiczigel and Lang, Flor et al.) is especially noteworthy when the samples for estimating Se and Sp are small which is usually the case for the validation studies of screening tests (Farnham et al., 2012).

Our aim was to present a new approach, a profile likelihood confidence interval (PLCI) for the interval estimation of the prevalence of a disease when Se and Sp of the diagnostic test are estimated from independent validation samples. We assessed its properties and compared its performance with the already available CIs proposed for the same problem, the ones by Lang and Reiczigel and Flor et al.

1.2 Method

A PLCI is obtained by inverting the likelihood ratio test (LRT). To give a short description of the LRT, let us assume that θ denotes a parameter of interest and ν some nuisance parameter (any of them may be vector-valued), and we are to test $H_0: \theta = \theta_0$ based on an observed sample x . The likelihood ratio Q is defined as the ratio of the following likelihoods: one maximized over the nuisance parameter ν while keeping θ at its null value θ_0 , and another maximized over both θ and ν . In formulas:

$$Q = \frac{L_0}{L_s}, \quad (3)$$

where:

$$L_0 = \max_{\nu} L(\theta_0, \nu, x), \quad (4)$$

and

$$L_s = \max_{\theta, \nu} L(\theta, \nu, x), \quad (5)$$

The test statistic of the LRT is

$$-2 \ln Q = -2 \ln \left(\frac{L_0}{L_s} \right). \quad (6)$$

Under appropriate regularity conditions, this test statistic has an approximate chi-square distribution on as many degrees of freedom as the dimension of θ . Thus, if θ is a scalar parameter, a $(1-\alpha)$ -level PLCI consists of those parameter values which are not rejected by the LRT at level α , that is:

$$\{\theta_0: -2 \ln Q < \chi_{1,1-\alpha}^2\} \quad (7)$$

where $\chi_{1,1-\alpha}^2$ is the $1-\alpha$ quantile of a χ^2 distribution on 1 degree of freedom (Millar, 2011).

In our case, prevalence of a disease (p) is the parameter of interest while Se and Sp of the diagnostic test are the nuisance parameters. Data comprise three independent samples of size n_{Se} , n_{Sp} , and n for the estimation of Se, Sp, and p , respectively. Let us denote by k_{Se} and k_{Sp} the number of correctly diagnosed out of n_{Se} and n_{Sp} , and k the number of positives out of n .

Under the assumption of fixed Se and Sp the distribution of k_{Se} , k_{Sp} , and k is binom(n_{Se} , Se), binom(n_{Sp} , Sp), and binom(n , \tilde{p}), where \tilde{p} is the unadjusted prevalence with Se and Sp being the true known diagnostic parameters and p is the true population prevalence. Thus, the likelihood function given the data n_{Se} , n_{Sp} , n , k_{Se} , k_{Sp} , and k looks like:

$$L(p, Se, Sp) = \binom{n_{Se}}{k_{Se}} Se^{k_{Se}} (1 - Se)^{n_{Se} - k_{Se}} \binom{n_{Sp}}{k_{Sp}} Sp^{k_{Sp}} (1 - Sp)^{n_{Sp} - k_{Sp}} \binom{n}{k} \tilde{p}^k (1 - \tilde{p})^{n - k}. \quad (8)$$

According to the above description, the $(1-\alpha)$ -level PLCI for p consists of those parameters p_0 , which are not rejected by the LRT at a significance level α , that is,

$$PLCI_{1-\alpha} = \left\{ p_0: -2 \ln \left(\frac{\max_{Se, Sp} L(p_0, Se, Sp)}{\max_{p, Se, Sp} L(p, Se, Sp)} \right) < \chi_{1,1-\alpha}^2 \right\}, \quad (9)$$

where $\chi_{1,1-\alpha}^2$ is the $1-\alpha$ quantile of a χ^2 distribution on 1 degree of freedom.

For the comparison of the CIs, we used simulated coverage probability (CP) and expected length (EL) as these are the most used features for the evaluation of intervals (Agresti & Coull, 1998; Gonçalves et al., 2012; Vollset, 1993). CP is the probability that the interval contains the true prevalence, preferred to be as close to the nominal level as possible (ideally always above the nominal level) while EL describes the intervals expected length:

$$EL(n, p) = \sum_{j=0}^n \binom{n}{j} p^j (1-p)^{n-j} (UCL(j) - LCL(j)), \quad (10)$$

where $UCL(j) - LCL(j)$ is the difference between the upper and lower limits of the corresponding interval. Note that both measures depend on the true parameters, so it is possible that for some parameter values one of the intervals, while for others another CI performs better.

During the evaluation of the proposed method, we observed that the coverage of the PLCI may be less than 90% at a nominal level of 95% if Se and Sp are above 99.5%. Fortunately, non-coverage occurred mostly when the observed values are extreme (0 or 100% observed prevalence, 100% observed Se or Sp). For example, if $n_{Se} = 100$, $n_{Sp} = 200$, $n = 300$, $k_{Se} = 95$, $k_{Sp} = 199$, $k = 0$, k is an extreme value. Coverage of PLCI can be improved considerably by applying a heuristic adjustment in such extreme cases. For the adjustments, one should calculate the CI endpoints, LCL and UCL also with the value next to the extreme and taking the arithmetical mean of this and the original endpoint, given this widens the interval. In the above example, we calculated the PLCI for $k = 0$ and also for $k = 1$ and took the average of the upper endpoints (the lower endpoint is 0 for both $k = 0$ and 1). In formula:

$$UCL_{\text{adjusted}} = (UCL_{k=0} + UCL_{k=1})/2. \quad (11)$$

If two or all three observed values (k , k_{Se} , and k_{Sp}) take extreme values, we consider the widest CI, that is, the union of the adjusted CIs.

The algorithm is implemented in an R function, using the optimization function `optim()` for maximization and interval halving for the determination of the CI endpoints. Likelihood is calculated on the logit scale, to avoid the necessity of optimization on bounded regions. The R function allows calculation of a one- or two-sided interval and specifying the required confidence level.

The function is public at: <https://github.com/Ragnar0ss/ProfileLikelihoodConfidenceInterval>.

For simplicity, the comparison of PLCI to the Lang-Reiczigel CI was made in the setting applied by Lang and Reiczigel in their respective paper, while comparison to the Flor interval was made in the setting applied by Flor et al. in their paper. The former comprises 5625 combinations of parameters. For true Se and Sp 1, 0.99, 0.95, 0.90, and 0.70 were used, while sample sizes for estimation of Se , Sp and prevalence were selected to be 30, 100, 300, 1000, and 3000. True prevalence was set to 0.005, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3, 0.5. Note that, due to symmetry of the CIs, considering prevalence values higher than 0.5 is not necessary (a CI for $p = p$, $Se = x$, $Sp = y$ and $p = 1-p$, $Se = y$, $Sp = x$ are mirror images of each other).

For each combination of the above listed parameters 20000 random sample triplets (for estimates of Se , Sp and prevalence) were generated and CP and EL were calculated. Note that the standard error of CP estimated from 20000 replications is 0.0015, providing sufficient precision of the simulation results. For length comparison we used the relative difference $100 \frac{EL_{PLCI} - EL_{Lang}}{EL_{Lang}} \%$, which is 0 when the two CIs have the same EL, and it is negative when PLCI is shorter.

The setting applied by Flor et al. uses a smaller data set with 1000 random combinations of n_{Se} , n_{Sp} , n , Se , Sp , p , where n_{Se} and n_{Sp} are chosen from the set of values 50, 100, 200, 500, 1000, 2000, 5000, n is a whole number chosen from [50, 2000]. Se and Sp are random numbers in [0.6, 1], and p is a random number in [0, 1]. For each parameter combination, 20000 samples were generated, and CP and EL were calculated from these samples. Detailed description of the Lang-Reiczigel and Flor et al intervals can be found in their respective papers.

1.3 Results and Discussion

1.3.1 Comparison of PLCI to the Lang-Reiczigel CI

Coverage probability of the adjusted PLCI is about the same as that of the Lang-Reiczigel CI: it reached the nominal level in 82% (Lang) and 89% (PLCI) of the studied 5625 cases and was less than 94% in as small as 12 and 27 times (0.21% and 0.48%), respectively. With PLCI this occurred when $n = 30$ and 100 (24 and 3 times) and with the Lang-Reiczigel CI when $n = 30$ and 3000 (3 and 9 times).

Coverage curves for $n_{Se} = n_{Sp} = n = 100$, with 90% and 99% Se and Sp , are shown in Figure 1. The curves show the advantage of adjustment of PLCI in case of high sensitivity and/or specificity: coverage probability without adjustment may fall below 92%, whereas with adjustment it remains above 94.4%. Coverage curves are depicted in the prevalence range from 0 to 0.5 because as discussed earlier the other half of the curves (for prevalence greater than 0.5) is a mirror-image of this with swapped sensitivity and specificity.

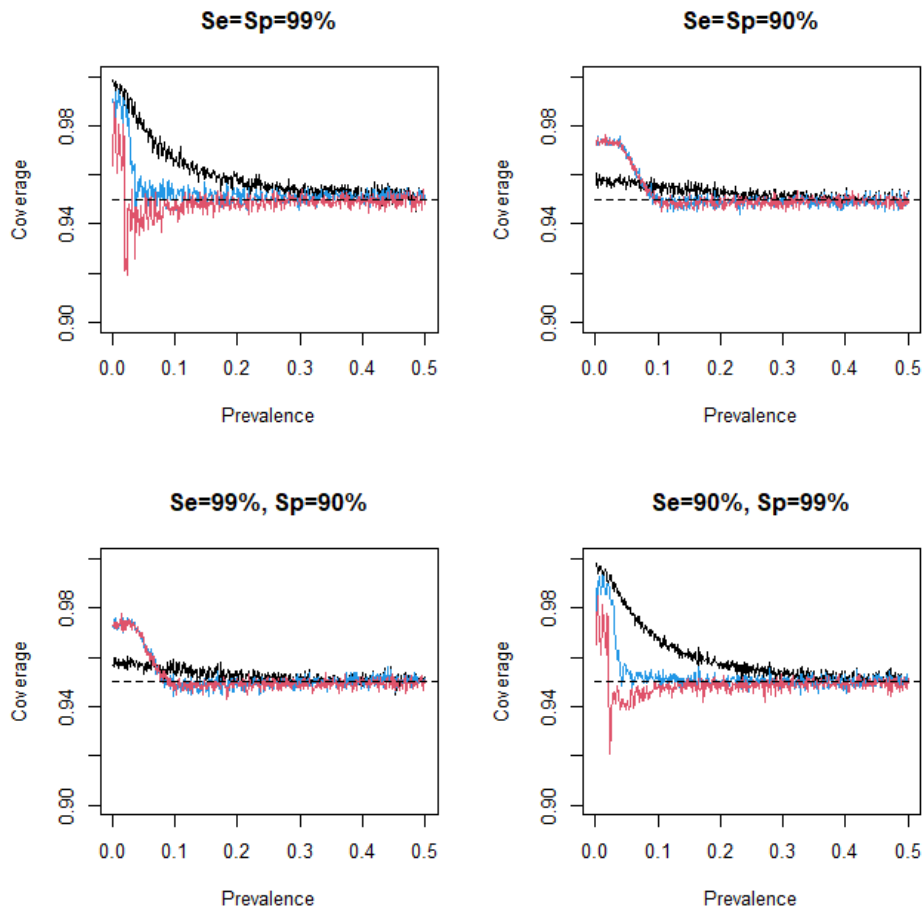


Figure 1. - Coverage curves of 95% CIs for $n_{Se} = n_{Sp} = n = 100$. Blue curve: PLCI with adjustment, red curve: PLCI without adjustment, black curve: Lang-Reiczigel CI.

Expected length of our new CI is on average smaller than that of the Lang-Reiczigel CI if the true prevalence is below 0.2 or above 0.8, while it is longer if true prevalence is between 0.3 and 0.7 (Table 1). Interestingly, the new CI also proved to be longer in the – practically rather seldom occurring – cases when sample sizes for sensitivity and specificity were 3000.

Table 1. - Relative difference between the expected length of 95% two-sided PLCI and Lang-Reiczigel CI by true prevalence (values for true prevalence > 0.5 follow from symmetry). Negative values indicate that PLCI is shorter (length of the Lang-Reiczigel CI is regarded as 100%).

	True prevalence					
	≤0.02	0.03	0.05	0.1	0.2	0.3-0.5
Relative length difference						
Overall	-7.8%	-5.5%	-4.0%	-2.3%	-0.0%	1.3%
If both Se and Sp ≥ 0.9	-9.2%	-6.2%	-4.4%	-2.6%	-0.1%	1.0%
If both Se and Sp ≥ 0.95	-11.0%	-7.4%	-5.1%	-3.2%	-0.2%	1.1%

Dependence of EL on Se and Sp is illustrated for sample sizes $n = 1000$ and $n_{Se} = n_{Sp} = 100$ in Table 2. Note that relative length difference in favor of the PLCI is in many cases more than 10%, while the difference in favor of the Lang-Reiczigel CI exceeds 3% only in two cases (namely when $Se = 0.7, Sp = 0.7, p = 0.5$, and $Se = 0.7, Sp = 1, p = 0.1$).

Table 2. - Length improvement of 95% two-sided PLCI compared to Lang-Reiczigel CI depending on Se, Sp and true prevalence for $n=1000, n_{Se} = n_{Sp} = 100$. Negative percentage values (with gray shading) indicate that PLCI is shorter (length of the Lang-Reiczigel CI is regarded as 100%). Darker shading indicates improvement greater than 5%.

Se	Sp	Prevalence								
		0.005	0.01	0.02	0.03	0.05	0.1	0.2	0.3	0.5
0.7	0.7	-9%	-9%	-9%	-9%	-8%	-6%	-1%	3%	8%
0.7	0.9	-3%	-4%	-4%	-4%	-3%	-1%	2%	3%	3%
0.7	0.95	-6%	-6%	-6%	-5%	-4%	-2%	1%	2%	2%
0.7	0.99	-21%	-17%	-13%	-10%	-4%	-1%	0%	1%	1%
0.7	1	-44%	-28%	-14%	-11%	-6%	4%	3%	3%	2%
0.9	0.7	-1%	-1%	-1%	-2%	-2%	-2%	0%	2%	3%
0.9	0.9	0%	-1%	-1%	-2%	-2%	-1%	0%	0%	1%
0.9	0.95	-5%	-5%	-5%	-4%	-3%	-2%	-1%	-1%	0%
0.9	0.99	-19%	-15%	-12%	-8%	-4%	-3%	-2%	-2%	-1%
0.9	1	-38%	-23%	-12%	-12%	-1%	2%	1%	1%	1%
0.95	0.7	1%	1%	0%	-1%	-1%	-2%	0%	1%	2%
0.95	0.9	1%	0%	-1%	-1%	-1%	-1%	-1%	0%	0%
0.95	0.95	-4%	-4%	-4%	-4%	-3%	-3%	-2%	-2%	-1%
0.95	0.99	-19%	-15%	-11%	-7%	-5%	-4%	-3%	-3%	-2%
0.95	1	-37%	-22%	-12%	-12%	0%	1%	0%	0%	-1%
0.99	0.7	3%	2%	1%	0%	0%	-1%	0%	1%	1%
0.99	0.9	2%	1%	0%	-1%	-1%	-1%	-1%	0%	-1%
0.99	0.95	-4%	-4%	-4%	-4%	-3%	-3%	-2%	-2%	-2%
0.99	0.99	-18%	-14%	-11%	-6%	-5%	-4%	-3%	-3%	-3%
0.99	1	-36%	-21%	-11%	-12%	0%	1%	0%	-1%	-1%
1	0.7	3%	3%	2%	1%	0%	0%	1%	2%	3%
1	0.9	2%	1%	0%	0%	-1%	-1%	0%	1%	1%
1	0.95	-3%	-3%	-3%	-3%	-3%	-2%	-1%	0%	0%
1	0.99	-18%	-14%	-10%	-6%	-4%	-3%	-2%	-2%	-1%
1	1	-35%	-20%	-11%	-11%	1%	1%	1%	1%	1%

1.3.2 Comparison of PLCI to the Flor interval

Comparison of the PLCI to the Flor interval was made on the same 1000 random combinations of n_{Se} , n_{Sp} , n , Se , Sp , p that was used by Flor et al. in their article with which they evaluated their CI. Coverage of the Flor interval is lower than that of PLCI and Lang-Reiczigel CI. Its minimum on this data set is as low as 88.4% at a nominal level of 95% (for PLCI and Lang CI it is 92.5% and 92.6%, respectively), and coverage probability falls below 93% in 7.7% of the studied 1000 cases (for PLCI and Lang CI it is 0.3% and 0.1%).

Length of the PLCI and Flor intervals are comparable to each other. Out of the 1000 parameter sets, 512 times the Flor interval while 488 times PLCI proved to be shorter. On average, Flor interval was 1% shorter but this shortness advantage was even smaller in case of higher Se and Sp . If both Se and Sp are at least 75%, although on average the Flor interval is still 0.6% shorter than the PLCI, in 54.8% of these cases PLCI is shorter.

As a summary, the proposed new CI has coverage probabilities comparable to those of the Lang-Reiczigel CI while its expected length is shorter, except when true prevalence is near 0.5. The Flor interval is comparable to the PLCI in terms of length but has much lower coverage probabilities.

1.3.3 Application examples

To illustrate the performance of the CIs in real applications, we applied them for the first three applications used in the article of Lang and Reiczigel (2014).

Example 1.

Boelaert et al. studied the prevalence of bovine herpesvirus-1 in cattle (Boelaert et al., 2000). In unvaccinated herds, 4060 out of 11,284 animals were found infected with the virus. Assuming that the Se of the gB-blocking ELISA is 0.99 (178/179 Kramps et al., 1994) and its Sp is 0.997 (358/359 De Wit et al., 1998), the true prevalence is 0.362. In this example, presumably because all three samples are quite large, all three methods deliver practically the same CI (PLCI: 0.349 to 0.371, Lang: 0.349 to 0.372, Flor: 0.348 to 0.372).

Example 2.

Anderson et al. studied the occurrence of trichomonad protozoa in free ranging songbirds (Anderson et al., 2009). Now we are using the prevalence of *Trichomonas gallinae* in house finches (*Carpodacus mexicanus*). Out of 2971 birds, 51 had the parasite, which resulted in an

apparent prevalence of 0.017. Se and Sp were found to be 0.97 (32/33), and 1 (20/20), respectively. Using these values, the point estimate for the true prevalence is 0.018, and the 95% Lang CI is 0 to 0.053. In this example both PLCI and Flor intervals are considerably shorter (0 to 0.023 and 0 to 0.019, respectively).

Example 3.

Faye et al. studied the prevalence of tuberculosis (TB) in dairy cattle in Uganda (Bernard et al., 2005). They applied a TB test on 11,862 animals and found an apparent prevalence of 0.06. The Se and Sp of the test was, according to Quirin et al., 0.80 (8/10), and 1 (12/12), respectively (Quirin et al., 2001). Using these data, a 95% Lang CI turns out to be 0 to 0.147, with a point estimate of .075. Here too, PLCI and Flor intervals are shorter, namely 0 to 0.116 and 0 to 0.090, respectively.

2 Confidence limits for risk differences and risk ratios adjusted for estimated sensitivity and specificity

The content of the chapter is based on my article with the same title, published in *Biostatistics & Epidemiology*, in 2023.

2.1 Background

Another important application of epidemiology is to evaluate how much of a specific disease is caused by a certain risk factor. Usually, to establish causal relationships and ultimately identify effective interventions, the occurrence of disease in a specific group of people exposed to a specific risk factor is compared to the disease occurrence observed in an unexposed group. This way, one can quantify the association between a risk (or protective) factor and a disease (or any other outcome). (Bailey et al., 2005)

Two of the most basic measures of association between diseases and risk factors are the risk ratio (RR) and risk difference (RD). RR is the ratio between the cumulative incidence in the exposed group and the cumulative incidence in the unexposed group, while RD can give information on how much greater the frequency of a disease is in the exposed group than in the unexposed group. (Gordis, 2013)

Methods to avoid bias are an important chapter of clinical and epidemiological research. Biases were often analysed and classified, like in the work of Sackett (Sackett, 1979) who identified 35 different types of biases in sampling and measurement, and evaluated their capacity to distort the point estimates of relative risks or odds ratios in epidemiology. The bias called “diagnostic suspicion bias” which affects mainly cohort studies is in the first place among measurement biases according to Sackett.

It is well known that in studies where the outcome is based on a diagnostic test, the estimate of the effect measures (RDs, RRs, or effect sizes in clinical trials) is unbiased only if Se and Sp are considered. However, it is less known that variance estimates may also be biased (and therefore the coverage of confidence intervals may be lower than the nominal) if Se and Sp are considered to be known, whereas they were estimated within trials with low sample size. Ignoring the random variability of Se and Sp estimates is not negligible.

As discussed in Section 1.1, an exact method for adjusting a single proportion (prevalence of disease) was presented by Reiczigel et al (Reiczigel et al., 2010) but this method was computationally very intensive, therefore an extension to RRs or RDs was not computationally

feasible. To our knowledge, some methods are available for the construction of confidence intervals for the RD and the RR if Se and Sp are regarded as known (Hahn et al., 2019; Lachenbruch, 1998; S.-F. Qiu et al., 2016, 2018; Reiczigel et al., 2017), but no methods have been published yet for interval construction when the diagnostic parameters are considered to be independent binomial estimates of the true Se and Sp of a certain diagnostic method. This consideration though may be important especially when Se and Sp of the test were estimated within studies with small sample sizes.

The method proposed in this section aims to estimate the CIs for RDs and RRs adjusted for estimated Se and Sp . The topic became recently especially relevant since the use of rapid antigen tests has gained widespread acceptance as an alternative method for diagnosis of COVID-19 outside of health care settings. In field studies aiming to estimate vaccine efficacy the adjustment for the variability of Se and Sp may be extremely important.

2.2 Method

Our approach combines two individual methods. One is the general method described by Zou and Donner (Zou & Donner, 2008) for calculating the CI for the difference between two effect measures based on their individual CI limits (and also applicable for the CI of ratio). The other is the method of Lang and Reiczigel (Lang & Reiczigel, 2014) to estimate the CI of a single proportion, adjusted for an estimated Se and Sp . Our method is computationally simple, and it can be applied also for the so-called differential classification errors as well (i.e. when the events in the two groups in which risks were estimated were detected with unequal sensitivities and specificities).

Assuming that there are two independent treatment arms in a clinical study, or two independent groups in an observational study, the confidence limits of risks in each group can be calculated by the method of Lang and Reiczigel, considering the variability of Se and Sp (i.e. the size of the samples from which Se and Sp were estimated). According to their method for the estimation of apparent prevalence (observed risk or observed response rate in a clinical study) $z_{crit}^2/2$ is added to each cell frequency, where z_{crit} denotes the percentile of the standard normal distribution corresponding to a certain (two-sided) confidence level. To improve the estimate of Se and Sp , 1 is added to each cell frequency (similar to Agresti's „add two success and two failures” method).

Given the three independent binomial estimates in each group: \hat{t}_i (cell frequency, response rate or observed risk), the observed sensitivity \widehat{Se}_i and the observed specificity \widehat{Sp}_i in both

groups ($i = 1, 2$ for differential misclassification), estimated based on samples of sizes n_i , $n_{Se,i}$ and $n_{Sp,i}$ respectively, the improved estimates \hat{t}'_i , \widehat{Se}'_i , \widehat{Sp}'_i are:

$$\hat{t}'_i = \frac{\hat{t}_i n_i + z_{crit}^2 / 2}{n_i + z_{crit}^2}, \quad (12)$$

$$\widehat{Se}'_i = \frac{\widehat{Se}_i n_{Se,i} + 1}{n_{Se,i} + 2}, \quad (13)$$

$$\widehat{Sp}'_i = \frac{\widehat{Sp}_i n_{Sp,i} + 1}{n_{Sp,i} + 2}. \quad (14)$$

The estimates used for the confidence interval constructions for the true risk in the populations are then:

$$\hat{R}_i = \frac{\hat{t}'_i + \widehat{Sp}'_{i-1}}{\widehat{Se}'_i + \widehat{Sp}'_{i-1}}. \quad (15)$$

In practice though, the risk estimates are truncated to the interval $[0, 1]$, in order to guarantee a valid proportion. Its confidence limits are given by:

$$\hat{R}_i + dR_i \pm z_{crit} \sqrt{var(R_i)}, \quad (16)$$

where

$$var(R_i) = \frac{\hat{t}'_i(1-\hat{t}'_i)/(n_i+z_{crit}^2) + \hat{R}_i^2 \widehat{Se}'_i(1-\widehat{Se}'_i)/(n_{Se,i}+2) + (1-\hat{R}_i)^2 \widehat{Sp}'_i(1-\widehat{Sp}'_i)/(n_{Sp,i}+2)}{(\widehat{Se}'_i + \widehat{Sp}'_{i-1})^2}, \quad (17)$$

and

$$dR_i = 2z_{crit}^2 \left[\hat{R}_i \frac{\widehat{Se}'_i(1-\widehat{Se}'_i)}{n_{Se,i}+2} - (1-\hat{R}_i) \frac{\widehat{Sp}'_i(1-\widehat{Sp}'_i)}{n_{Sp,i}+2} \right]. \quad (18)$$

Applying the above method for both groups in case of differential misclassification (which are assumed to be independent), the adjusted risk estimates \hat{R}_1 and \hat{R}_2 , and their confidence limits (l_1, u_1) , (l_2, u_2) are obtained.

Once the risk estimates \hat{R}_1 and \hat{R}_2 , and their confidence limits (l_1, u_1) , (l_2, u_2) are computed as described above, the method of Zou and Donner can be applied to estimate the risk difference $\widehat{RD} = \hat{R}_1 - \hat{R}_2$ and the risk ratio $\widehat{RR} = \hat{R}_1 / \hat{R}_2$ between the two groups, along with their confidence intervals of the same confidence level.

Zou and Donner presented a general method to construct a confidence interval for a difference between two independent parameter estimates, needing as a prerequisite only the two separate confidence intervals of the two parameter estimates.

If (l_1, u_1) and (l_2, u_2) are the confidence intervals of a certain level for the parameters θ_1 and θ_2 with estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, then the confidence limits (LCL, UCL) for the difference $\hat{\theta}_1 - \hat{\theta}_2$ (having the same confidence level as the two independent confidence intervals) can be computed with the formulae:

$$LCL = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2}, \quad (19)$$

$$UCL = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(\hat{\theta}_2 - l_2)^2 + (u_1 - \hat{\theta}_1)^2}. \quad (20)$$

This approach can be applied for RDs for which the estimates and their confidence intervals are adjusted for the Se and Sp of the diagnostic methods (even in the case when the two risks are estimated applying diagnostic methods with different Se and Sp).

The same approach can also be used to estimate the confidence interval of RRs if the above formulae are applied on a logarithmic scale, and results are then transformed back by exponentiation:

$$LCL = \ln(\hat{\theta}_1) - \ln(\hat{\theta}_2) - \sqrt{(\ln(\hat{\theta}_1) - \ln(l_1))^2 + (\ln(u_2) - \ln(\hat{\theta}_2))^2}, \quad (21)$$

$$UCL = \ln(\hat{\theta}_1) - \ln(\hat{\theta}_2) + \sqrt{(\ln(\hat{\theta}_2) - \ln(l_2))^2 + (\ln(u_1) - \ln(\hat{\theta}_1))^2}. \quad (22)$$

As recommended in the original paper, for extreme cases for which logarithm is not defined, $\hat{\theta}_i$ and l_i are set to $1/2n_i$ if they are equal to 0, and $\hat{\theta}_i$ and u_i are set to $1 - (1/2n_i)$ if they are equal to 1. Also, in case if the denominator is 0 for the point estimates of the RR the term $1/2n_i$ is added, where n_i is the size of sample l from which the risk in the denominator was estimated.

The method of Zou and Donner was originally developed for two independent binomial variables. However, when applying a Rogan-Gladen-type transformation to the two independent estimates \hat{t}_1 and \hat{t}_2 based on observed data, the transformed values obtained by equation (15) remain independent only in case of differential misclassification errors. Otherwise, in case of non-differential misclassification, they will be slightly correlated because both will depend on the same \widehat{Se} and \widehat{Sp} estimates. However, the parameter estimates \hat{R}_1 and \hat{R}_2 will be positively correlated, and according to the exact type I error rates and coverage rates, it can be concluded that the influence of this small correlation on the method performance is minimal.

2.3 Discussion

2.3.1 Evaluation of the method performance

The exact coverage probabilities (*CP*) for confidence intervals of RDs and RRs calculated by the proposed new method were computed by evaluating all the possible (n_1+1) , (n_2+1) , $(n_{Se_1}+1)$, $(n_{Sp_1}+1)$, $(n_{Se_2}+1)$ and $(n_{Sp_2}+1)$ outcomes for the given true R_1 , R_2 , Se_1 , Se_2 , Sp_1 , Sp_2 parameters:

$$\begin{aligned}
 CP = & \sum_{r_1=0}^{n_1} \sum_{r_2=0}^{n_2} \sum_{r_{Se_1}=0}^{n_{Se_1}} \sum_{r_{Sp_1}=0}^{n_{Sp_1}} \sum_{r_{Se_2}=0}^{n_{Se_2}} \sum_{r_{Sp_2}=0}^{n_{Sp_2}} \binom{n_1}{r_1} R_{Obs_1}^{r_1} (1 - R_{Obs_1})^{n_1-r_1} \\
 & \binom{n_2}{r_2} R_{Obs_2}^{r_2} (1 - R_{Obs_2})^{n_2-r_2} \binom{n_{Se_1}}{r_{Se_1}} Se_{true_1}^{r_{Se_1}} (1 - Se_{true_1})^{n_{Se_1}-r_{Se_1}} \binom{n_{Sp_1}}{r_{Sp_1}} Sp_{true_1}^{r_{Sp_1}} (1 \\
 & - Sp_{true_1})^{n_{Sp_1}-r_{Sp_1}} \\
 & \binom{n_{Se_2}}{r_{Se_2}} Se_{true_2}^{r_{Se_2}} (1 - Se_{true_2})^{n_{Se_2}-r_{Se_2}} \binom{n_{Sp_2}}{r_{Sp_2}} Sp_{true_2}^{r_{Sp_2}} (1 - Sp_{true_2})^{n_{Sp_2}-r_{Sp_2}} I_{[LCL,UCL]}(RD_{true}), \quad (23)
 \end{aligned}$$

where r_1 and r_2 are the number of positive outcomes observed in the two samples, R_{Obs_1} and R_{Obs_2} are the observed (unadjusted) risks, r_{Se_1} and r_{Sp_1} are the number of true positive and true negative cases respectively, based on which sensitivity and specificity were estimated for diagnostic test 1 applied in sample 1, whereas r_{Se_2} and r_{Sp_2} are the number of true positive and true negative cases based on which *Se* and *Sp* were estimated for diagnostic test 2 applied in sample 2.

$I_{[LCL,UCL]}(RD_{true})$ as an indicator value equals to 1 if the interval $[LCL, UCL]$ contains the true value of the risk difference and 0 otherwise. A similar coverage probability can be computed for risk ratios, replacing the indicator function in the previous formula with $I_{[LCL,UCL]}(RR_{true})$.

The exact type 1 error rate of the inference based on confidence intervals for risk ratios was computed by evaluating all the possible (n_1+1) , (n_2+1) , $(n_{Se_1}+1)$, $(n_{Sp_1}+1)$, $(n_{Se_2}+1)$ and $(n_{Sp_2}+1)$ outcomes for true $R_1=R_2$, and given true Se_1 , Se_2 , Sp_1 , Sp_2 values.

$$\begin{aligned}
 \text{type I error} = & \sum_{r_1=0}^{n_1} \sum_{r_2=0}^{n_2} \sum_{r_{Se_1}=0}^{n_{Se_1}} \sum_{r_{Sp_1}=0}^{n_{Sp_1}} \sum_{r_{Se_2}=0}^{n_{Se_2}} \sum_{r_{Sp_2}=0}^{n_{Sp_2}} \binom{n_1}{r_1} R_{Obs_1}^{r_1} (1 - R_{Obs_1})^{n_1-r_1} \\
 & \binom{n_2}{r_2} R_{Obs_2}^{r_2} (1 - R_{Obs_2})^{n_2-r_2} \binom{n_{Se_1}}{r_{Se_1}} Se_{true_1}^{r_{Se_1}} (1 - Se_{true_1})^{n_{Se_1}-r_{Se_1}} \binom{n_{Sp_1}}{r_{Sp_1}} Sp_{true_1}^{r_{Sp_1}} (1 \\
 & - Sp_{true_1})^{n_{Sp_1}-r_{Sp_1}}
 \end{aligned}$$

$$\binom{n_{Se_2}}{r_{Se_2}} Se_{true_2}^{r_{Se_2}} (1 - Se_{true_2})^{n_{Se_2} - r_{Se_2}} \binom{n_{Sp_2}}{r_{Sp_2}} Sp_{true_2}^{r_{Sp_2}} (1 - Sp_{true_2})^{n_{Sp_2} - r_{Sp_2}} I(1 \notin [LCL, UCL]), \quad (24)$$

where all the notations are as described above for the coverage rate.

The exact type 1 error rate of the inference based on confidence intervals for risk differences can also be computed with a slightly different indicator value $I(0 \notin [LCL, UCL])$, where all the notations are as described above for the coverage rate.

Figure 2 illustrates the exact coverage rates for the confidence interval for the risk ratios, while figure 3 illustrates the same for the risk difference.

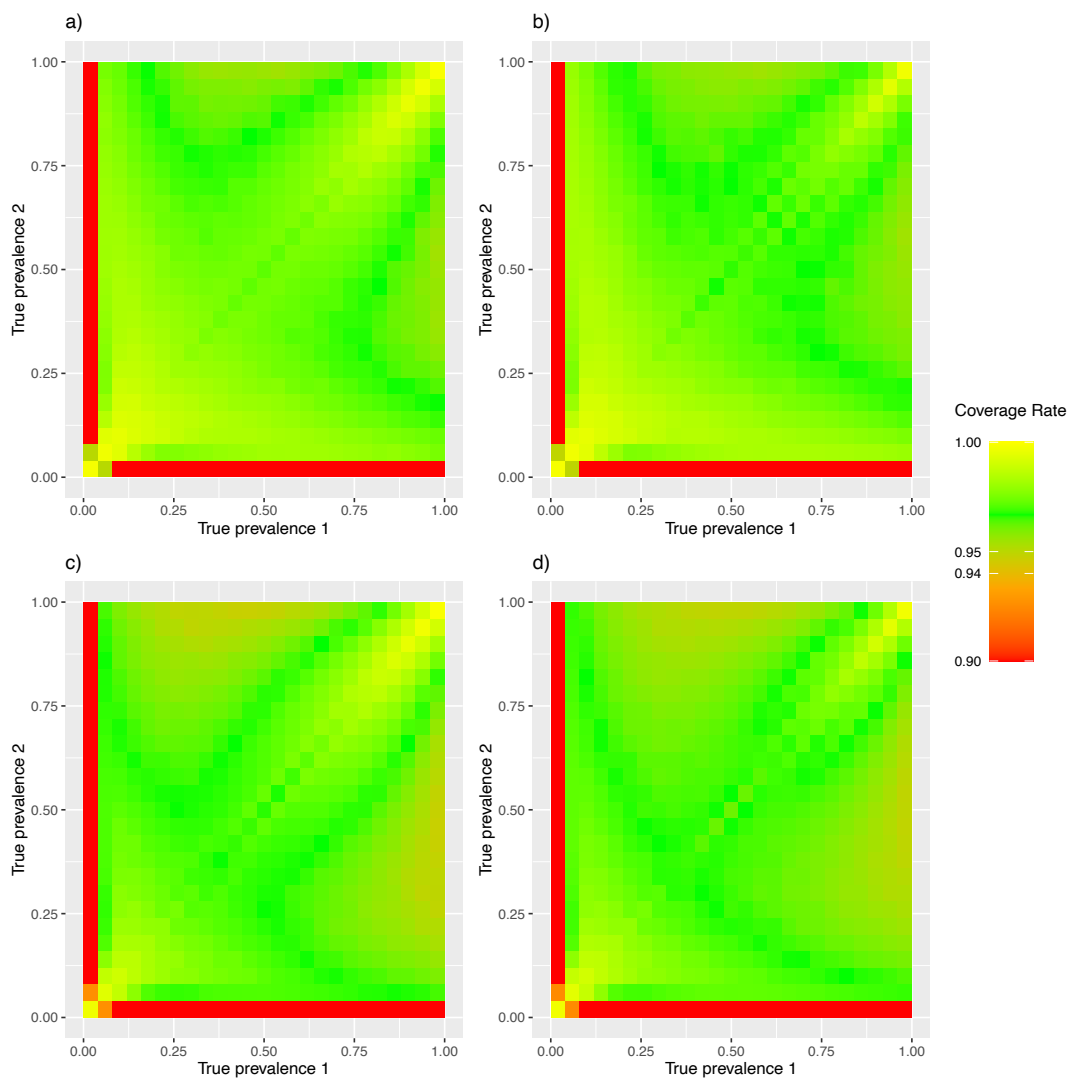


Figure 2. - Exact coverage rates for the 95% confidence interval of the RR with different sample sizes used for the estimation of sensitivity and specificity. a) $n_{Se} = 20, n_{Sp} = 20$. b) $n_{Se} = 50, n_{Sp} = 20$. c) $n_{Se} = 20, n_{Sp} = 50$ and d) $n_{Se} = 50, n_{Sp} = 50$. True Se and Sp are 0.9 in both groups. n_1 and n_2 are both selected to be 20.

Increasing the size of the sample used to estimate sensitivity and specificity results in a coverage rate of the confidence interval closer to the nominal level (95%), most probably due to the more precise estimation of the true prevalence (or response rate in a clinical trial setting). The exact coverage probabilities are apparently better for the interval of the risk ratio as in that case poor coverage is experienced only when one of the true prevalence lies at the margins of the parameter space. For the interval of the RD poor coverage was observed near the upper left corner, when high ($p_2 > 0.50$) true prevalence values were deducted from relatively small ($p_1 < 0.50$) ones.

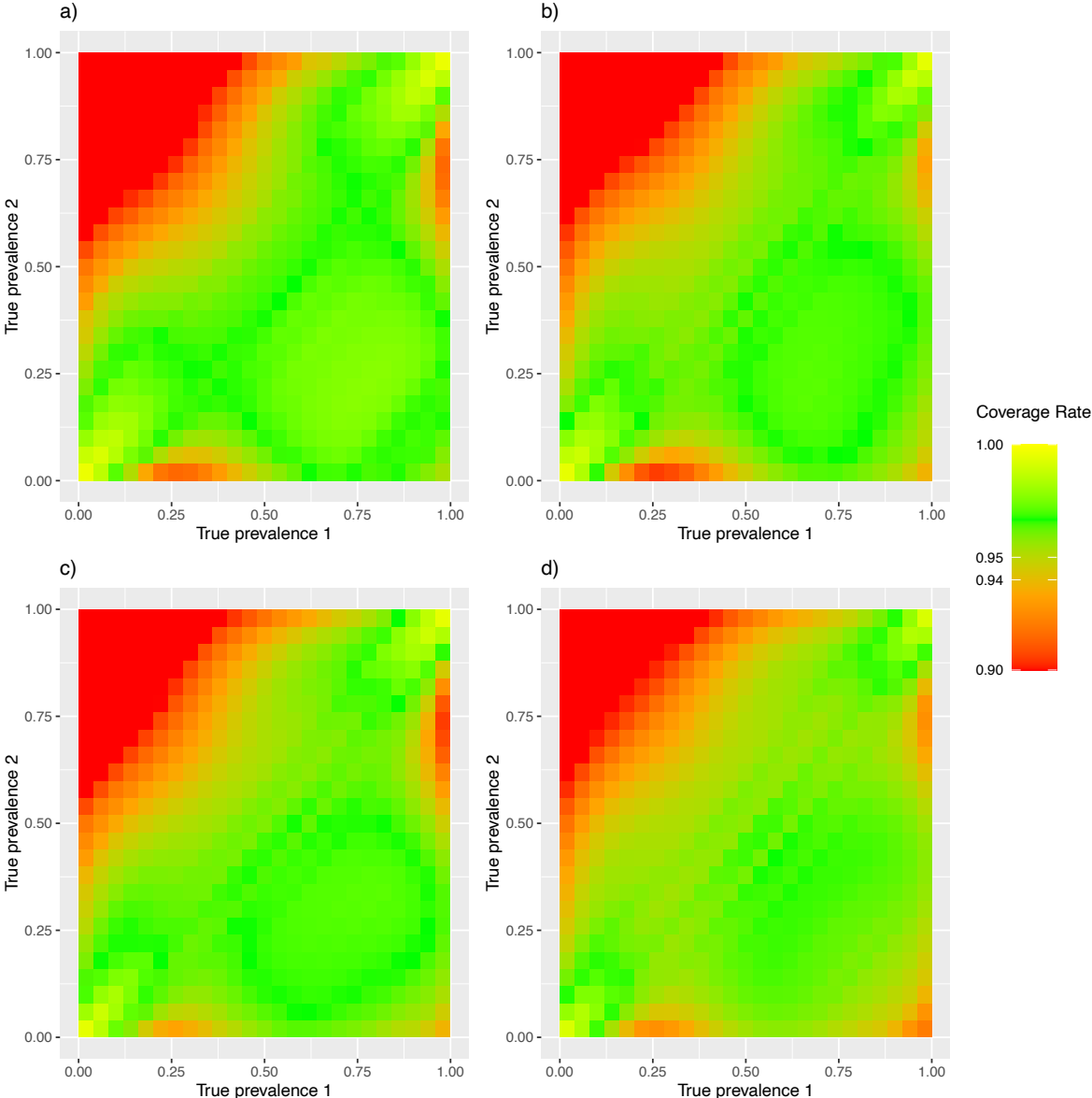


Figure 3. - Exact coverage rates for the 95% confidence interval of the RD with different sample sizes used for the estimation of sensitivity and specificity. a) $n_{Se} = 20, n_{Sp} = 20$. b) $n_{Se} = 50, n_{Sp} = 20$. c) $n_{Se} = 20, n_{Sp} = 50$ and d) $n_{Se} = 50, n_{Sp} = 50$. True Se and Sp are 0.9 in both groups. n_1 and n_2 are both selected to be 20.

Figure 4 illustrates the dependence of coverage rates for the confidence interval for the RD on the true sensitivity and specificity values. For both measures, the closer the true sensitivity and specificity is to 1, the better the coverage probabilities are. Only the plots for the risk difference are included for illustration because figures were similar for risk ratios.

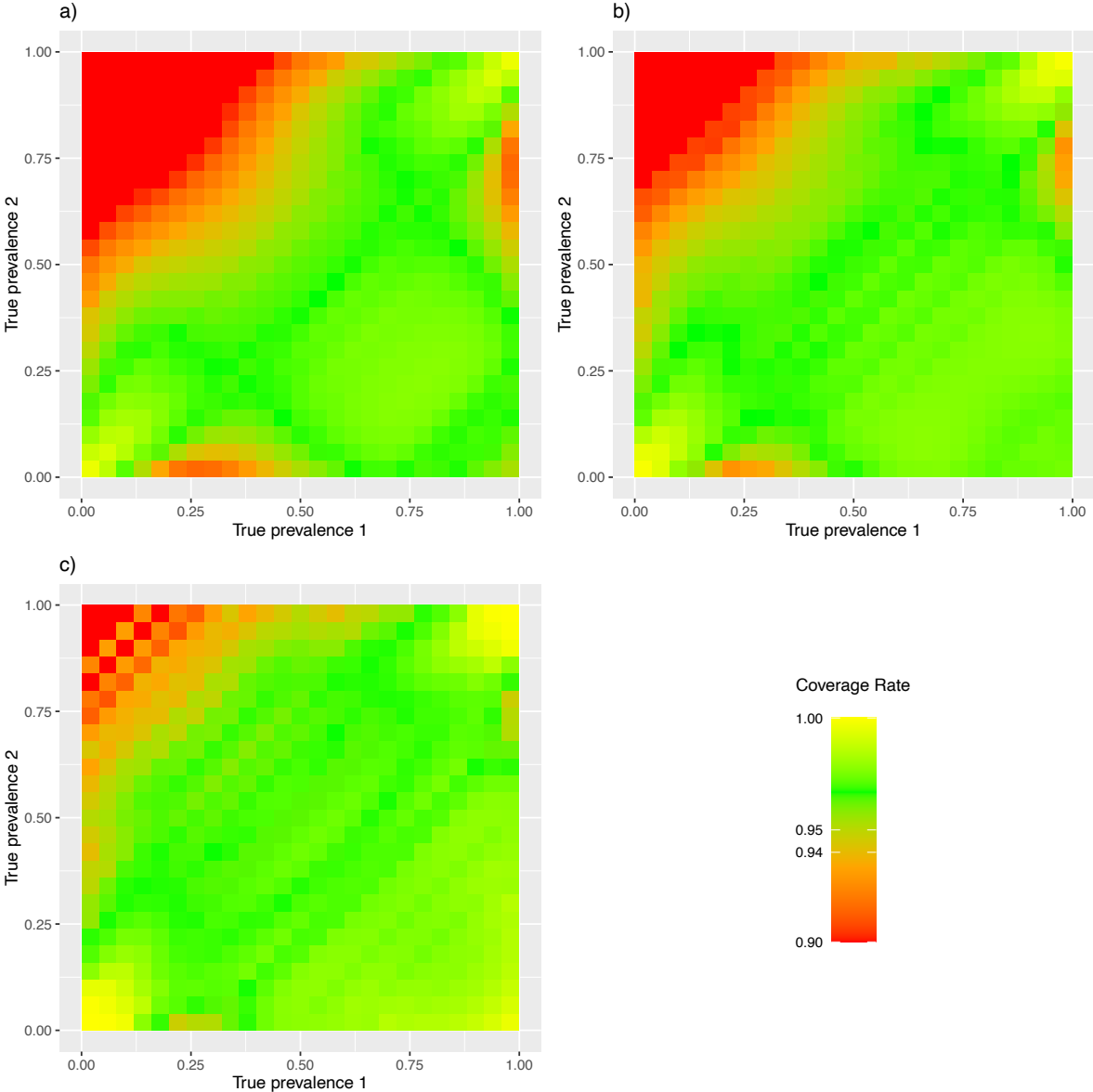


Figure 4. - Exact coverage rates for the 95% confidence interval of the RD with different true sensitivity and specificity values. a) $Se = Sp = 0.90$. b) $Se = Sp = 0.95$. c) $Se = Sp = 0.99$. $n_{Se} = 20$, $n_{Sp} = 20$ and n_1 and n_2 are both selected to be 20.

The exact type 1 error rates for inferences based on confidence intervals can easily be computed from the exact coverage rates ($1 - \text{exact coverage rate}$). For both inferences (based on RD and RR confidence intervals), the increase of the sample size or the increase of the true Se and Sp results in an alpha level closer to nominal 5%.

To illustrate some additional properties of our new method, the influence of sample size from which sensitivity and specificity were estimated on the width of confidence intervals calculated by the proposed method was assessed. Table 3 presents a constructed example with observed RR=1 (observed response rates in the two arms are both 30/60 and 40/80, respectively), Se estimate of 0.95, Sp estimate of 0.9, and the sample size for Se and Sp estimation incremented by steps of 30. As expected, the width of CI was decreasing when the sample sizes increased, converging to the width of CI calculated ignoring the random variability of Se and Sp.

Table 3. - Adjusted 95% CIs of RR for different n_{Se} and n_{Sp} values ($n_{Se}=n_{Sp}$), the observed response rates in the two arms are 30/60 and 40/80, respectively

n_{Se}	n_{Sp}	95% CI average width
30	30	0.01001
60	60	0.00784
90	90	0.00738
120	120	0.00716
150	150	0.00704
180	180	0.00696
210	210	0.00691
240	240	0.00687
270	270	0.00683
300	300	0.00681

The sample sizes n_{Se} and n_{Sp} do not seem to influence the width of the CI to the same extent for the observed response rates in the two arms. In Table 4 the 95% CI width of RR is presented for different n_{Se} and n_{Sp} pairs (the observed responses are the same as described for Table 3). In Table 4, $n_{Se} + n_{Sp}$ is constant within each row, but the CI width is larger when n_{Se} is larger than n_{Sp} , meaning that the variability in specificity influences on a greater extent the CI width for the selected parameter values. A similar phenomenon occurs in case of RDs, as presented in Table 5 (observed responses the same as in Table 3).

Table 4. - Adjusted 95% CIs of RR for different n_{Se} and n_{Sp} values ($n_{Se} \neq n_{Sp}$), the observed response rates in the two arms are 30/60 and 40/80, respectively

n_{Se}	n_{Sp}	95% CI mean width	n_{Se}	n_{Sp}	95% CI mean width
90	30	0.00990	30	90	0.00764
120	60	0.00777	60	120	0.00725
150	90	0.00734	90	150	0.00708
180	120	0.00714	120	180	0.00699
210	150	0.00703	150	210	0.00692
240	180	0.00695	180	240	0.00688

Table 5. - Adjusted 95% CIs of RD for different n_{Se} and n_{Sp} values ($n_{Se} \neq n_{Sp}$), the observed response rates in the two arms are 30/60 and 40/80, respectively

n_{Se}	n_{Sp}	95% CI mean width	n_{Se}	n_{Sp}	95% CI mean width
90	30	0.00358	30	90	0.00343
120	60	0.00326	60	120	0.00320
150	90	0.00315	90	150	0.00312
180	120	0.00310	120	180	0.00308
210	150	0.00307	150	210	0.00306
240	180	0.00308	180	240	0.00304

Our method is easy to use and computationally feasible. Its performance is better in case of differential misclassification than in case of a non-differential one, but still the coverage probabilities are close to the nominal confidence level except when at least one of the prevalences (or one of the response rates in a clinical trial) is close to 0. In those cases, the coverage rates are much below the nominal confidence level and the method should not be applied.

There are no universal cut-off values above which the influence of sample sizes from which sensitivity and specificity were estimated can be ignored. However, Table 3 indicates that the cut-off value of $n = 300$ is a good indicator, below it the adjustment for the variability of

sensitivity and specificity should not be ignored, and the method presented in this paper is recommended to be applied.

2.3.2 Application examples

Example 1.

Qiu et al (S. B. Qiu et al., 2012) studied the occurrence of *Trichomonas gallinae* in domestic pigeons, in subtropical southern China. A total of 319, apparently healthy domestic pigeons from 7 farms were tested by culturing the parasites, as described by Anderson et al (Anderson et al., 2009). Sensitivity and specificity were not taken into account in this study, however, from the reference paper of Anderson et al it turns out that sensitivity and specificity were 0.97 (32/33) and 1 (20/20), respectively. Level of the management and environmental sanitation (LMES) was studied as a risk factor of infection. LMES from different farms was classified as High/Low. At farms with high level LMES 18 infected pigeons were found among the total of 67 tested, whereas at farms with low level LMES 90 infected pigeons were found among the total of 252 tested. The unadjusted risk ratio (high versus low LMES) and its 95% CI was 0.752 (0.490 to 1.155). Since the specificity of the diagnostic method was 1 (and it was considered to be independent of LMES classification), the point estimate of the risk ratio was not influenced by adjustment for sensitivity and specificity. However, the width of the 95% confidence intervals after adjustment reflected the additional uncertainty of an imperfect classification. Adjusting for sensitivity and specificity, considering them as known parameters with the combination of the methods of Reiczigel (adjusting the lower and upper endpoints of the Wilson interval with known sensitivity and specificity) and Zou (for the calculation of the confidence interval for the RR) (Reiczigel et al., 2010; Zou & Donner, 2008), the CI was (0.482 to 1.122). Adjusting for sensitivity and specificity, considering them as random variables from samples of size 33 and 20, respectively, the CI was (0.321 to 1.276). Thus, adjusting for an estimated sensitivity and specificity led to an increase in CI width of 49.2% compared to the CI corrected for known sensitivity and specificity, and an increase of 43.6% compared to the uncorrected CI. This increase was large in spite of the fact that both Se and Sp estimates were very high for this diagnostic test, and it is mainly due to the low size of the samples from which sensitivity and specificity were estimated.

Example 2.

A randomized controlled field trial was conducted in Colima, Mexico to investigate the effect of a special intervention consisted of a community participation program focused on the ecosystem for the prevention of dengue fever (Newton-Sánchez et al., 2020). The control group received the usual official prevention programs. The incidence of dengue was estimated

in both groups with the appearance of the de novo IgM antibodies based on the results of a rapid Dengue-Duo (PanBio®) immunochromatographic test. On the official website of the manufacturer the characteristics of the test is available (*Panbio™ DENGUE DUO IgM CAPTURE AND IgG CAPTURE ELISA Test Specifications*, n.d.) with a primary serological sensitivity of 94.7% (54/57) and serological specificity of 100% (83/83). The incidence of dengue fever in the interventional group (community program focused on the ecosystem) was 15.53% (127/818) while in the control group (usual prevention) 13.58% (135/994) with an unadjusted risk ratio and 95% CI of 1.14 (0.91 – 1.43). Adjusting the risk ratio for sensitivity and specificity as constants the risk ratio and CI was 1.21 (0.914 to 1.429), while adjusting the risk ratio for sensitivity and specificity as random variables from sample sizes 57 and 83 gave the CI (0.827 to 1.633). The adjustment for an estimated sensitivity and specificity led to an increase in CI width of 56.5% compared to the CI corrected for known sensitivity and specificity, and a slightly less increase of 56.1% compared to the unadjusted CI.

Example 3.

Guidugli et al evaluated the effectiveness and safety of the antibiotic doxycycline for the prophylaxis of leptospirosis (Guidugli et al., 2000). The diagnosis of the participants was based on the microscopic agglutination test with a reported sensitivity of 55.3% (93/168) and specificity of 95.8% (162/169) (Niloofoa et al., 2015). The prevalence of leptospirosis in the group treated with doxycycline was 0.6% (3/509) while the prevalence in the placebo group was 4.87% (25/513). The unadjusted point estimate of the risk difference (placebo group – treatment group) with the corresponding 95% CI calculated by the Wald method (Newcombe, 1998) was 4.284 percentage points (2.306 to 6.262). The adjusted 95% CI using the adjustments considering sensitivity and specificity as constants gave the point estimate and CI 8.30 (4.619 to 12.768), whereas the adjustments for sensitivity and specificity considered as random variables resulted in the 95% CI (-8.018 to 5.296). As expected, due to the poor sensitivity of the diagnostic method, the adjustment had a big impact on the estimates: adjustment for random sensitivity and specificity led to an increase in CI width of 236.6%, whereas adjustment for known sensitivity and specificity resulted in a smaller increase of 106% compared to the uncorrected CI.

Example 4.

A total of 200 women were observed in a prospective case control study to compare the seroprevalence of *Toxoplasma gondii* in pregnant women in Colombo, Sri Lanka (Subasinghe et al., 2011). The test group was 100 women with a history of a spontaneous miscarriage within 28 weeks of pregnancy and who presented within six months of the event and the control group consisted of 100 healthy pregnant women within 28 weeks of pregnancy having no medical

complications. The seroprevalence of *Toxoplasma gondii* was 28% (28/100) in the control group while it was 17% (17/100) in the test group, resulting in the unadjusted risk difference (control group – test group) point estimate of 11 percentage points with the Wald 95% CI of (-0.4737 to 22.474). For the confirmation of the disease status OnSite™ Toxo IgG/IgM Rapid Tests were used with a sensitivity of 91.67% (22/24) and specificity of 99% (297/300) (*Instruction for Use for the OnSite Toxo igG/igM Rapid Test*, n.d.). The adjusted 95% CI for known constant sensitivity and specificity yielded a 95% CI of (-0.676 to 24.577), and if sensitivity and specificity were considered as random variables the 95% CI of the risk difference was (-1.738 to 26.776). The adjustment for an estimated sensitivity and specificity led to an increase in CI width of 12.91% compared to the CI corrected for known (constant) sensitivity and specificity, and an increase of 24.26% compared to the uncorrected CI.

3 The effect of misclassification on sample size for one and two-sample tests with binary endpoints

The content of the chapter is based on my article with the same title, published in Journal of Biopharmaceutical Statistics, in 2024.

3.1 Background

Testing for binary data with one- and two-sample statistical tests are extremely common in epidemiology and medical statistics (Bland, 2015). One-sample tests can be used among others when assessing freedom from disease or approving diagnostic tests, and in industrial quality control, evaluation of medical devices, and in clinical trials of rare diseases (Cameron & Baldock, 1998; Cheng & Zhen, 2021; Feld et al., 2015; Khan, Sarker, & Hackshaw, 2012; Lu, Li, & Xu, 2020). The left-sided alternative ($H_0: p = p_0$ against $H_a: p < p_0$) is applied for example in proving freedom from disease, while the right-sided one ($H_0: p = p_0$ against $H_a: p > p_0$) is used if one wants to prove that a particular exposure increases the probability of getting a disease. Two-sample tests ($H_0: p_1 = p_2$ against $H_a: p_1 \neq p_2$ for the two-sided scenario) are even more frequent in medical research and can be used to compare different groups on a binary response variable, for example to compare the prevalence of a disease between two populations (Agresti, 2012).

In many cases, the outcomes may be wrongly classified, for example in a clinical scenario when we are comparing the prevalence of a disease between two groups, determined by a diagnostic rapid test. Usually, a diagnostic test has less than 100% Se and Sp which can be accounted for at the design and at the analysis of a study. The main issue with ignoring misclassifications is that it can lead to a biased result. Depending on whether we decide to handle the misclassifications and if we do so, when, four different scenarios must be treated and discussed separately.

If misclassifications are accounted for both during the planning and the analysis of a study, then it will have no effect on statistical power, and the result will not be biased. If no correction is applied either during the planning or the analysis phase, the power of the test will remain as planned, but the result will reflect the apparent prevalence rather than the true one. There is only a theoretical possibility of considering misclassifications only at the planning stage, but in practice, this is rare. If someone designs a study while accounting for misclassifications during sample size calculation, it is unlikely that they would forget to correct for them during the

analysis. In this chapter, we examined the single group where misclassifications were ignored in the planning phase but corrected for during the analysis.

In various test scenarios and alternatives, we investigated what happens if the sample size is determined based on the power calculated during the planning phase, but the analysis at the end of the study adjusts for misclassifications using the parameters of the diagnostic test. In such cases, the power (i.e., the probability of rejection) can decrease significantly.

To illustrate the effect of misclassification, Figure 5 shows how the power of the one-sample binomial test depends on sensitivity and specificity for $p_0 = 0.02$, $p_a = 0.05$, two-sided alternative, and sample size 400. The power decreases considerably when either of the diagnostic parameter values are below 1. The effect of Sp is clearly more apparent in this particular example which is due to the parameter values p_0 and p_a tested. As a rule, if the true prevalence is low (near 0) the power of the test is influenced more strongly by the Sp of the diagnostic test, whereas if the true prevalence is high (near 1) the power is influenced more strongly by Se .

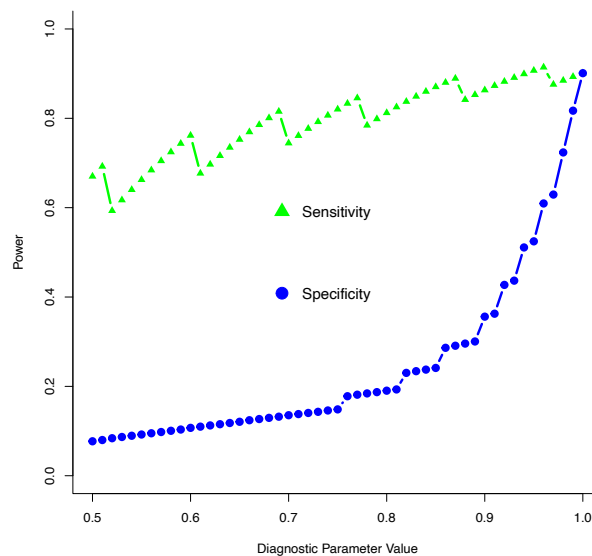


Figure 5. - Power of the exact binomial test as the function of the Se or Sp , while the other parameter is fixed at 1 ($p_0 = 0.02$, $p_a = 0.05$, alternative = "two-sided", $n = 400$)

The sample size required for a statistical test is mainly driven by the will to reach a particular power. The sample size needed for the same power in a scenario where there is an opportunity for misclassification is higher than it would be without misclassification. There are analysis methods accounting for misclassification (Lang & Reiczigel, 2014; Reiczigel, Földi, & Ózsvári, 2010; Hársfalvi & Reiczigel, 2023, Hársfalvi & Singer, 2023) but these cannot help if at the

design stage of the study the power loss caused by the potential misclassification of the used diagnostic test is not compensated with a justified increase of the sample size. Ignoring the possibility of misclassification in the sample size calculation may easily result in an underpowered, inconclusive study, causing considerable financial loss and raising ethical concerns.

To calculate the sample size for the one-sample test we need the prescribed alpha level and power, the null proportion p_0 (that in the null hypothesis) and the assumed true proportion p_a for which the prescribed power should be reached. For the two-sample scenario, in addition to the prescribed alpha and power we need the assumed true proportions in each group, p_1 and p_2 (Chow, Shao, & Wang, 2008; Suresh & Chandrashekara, 2012).

Most books on sample size calculation do not mention misclassification at all. Others have a short note declaring this as a problem advised to account for but none of them have clear instructions for researchers. (Chow et al., 2008; Julious, 2009; Kieser, 2020; Ryan, 2013). Intuition may suggest that if the probability of misclassification is as low as a few percent in both directions, the increase in sample size is ignorable, but this is not true. To illustrate this, we carried out extensive simulations. Furthermore, we developed functions for sample size calculation for both the one-sample and two-sample proportion tests under misclassification and investigate how the necessary sample size depends on different parameters.

Two different approaches can be used for the sample size calculations in the presence of misclassifications. The first approach (less applicable in practice) is when both Se and Sp of the diagnostic test are considered as fix, known parameters, available at the design stage of a study. The second, more realistic approach is to take the uncertainty in both Se and Sp into account and consider them as estimated quantities, acknowledging that these are also prone to errors. The real-world study example for the first approach is when one takes the Se and Sp of the diagnostic test into account during the design, but only as fixed parameters, and the second approach is when not only Se and Sp are used in the sample size calculations but also the size of the samples used for the estimation of these parameters, as the metric of their uncertainty. In case one or both misclassification parameters are completely unknown, straight forward sample size calculation is not possible. In these scenarios sensitivity analysis can be conducted to better understand the effect of the different possible misclassification parameter values on the required sample size.

We developed R functions for the sample size calculations under both approaches, while using multiple asymptotic (the Wald-test, the Wilson's score test and the Agresti-Coull-test) and

exact one-sample tests for proportions (the Clopper-Pearson and the Blaker test) complemented by one specific case of the two-sample test capable of handling misclassification suggested by Hársfalvi and Singer (Hársfalvi & Singer, 2023).

3.2 Methods

3.2.1 Introduction of the test methods

3.2.1.1 One-sample tests

For the one-sample tests we used the exact binomial test (a.k.a. the Clopper and Pearson test) (Chow et al., 2008; Clopper & Pearson, 1934) and the asymptotic test suggested by Wilson (Wilson, 1927). For a short description of these two, let X denote a variable from a binomial (n, p) distribution and x its observed value. Let n be the sample size and $b_{n,p}(x) = P(X = x)$ denote the probability mass function of the binomial distribution with parameters p and n . Assume that based on the observation of x we test for $H_0: p = p_0$ against $H_a: p \neq p_0$ (two-tailed test) or $H_a: p < p_0$ or $H_a: p > p_0$ (one-tailed tests). The alpha level critical region of the test, calculated by inverting the Clopper-Pearson exact confidence interval, is as follows:

$$\{x: \sum_{i=0}^x b_{n,p}(i) \leq \alpha\} \quad (\text{left-tailed test}) \quad (25)$$

$$\{x: \sum_{i=x}^n b_{n,p}(i) \leq \alpha\} \quad (\text{right-tailed test}) \quad (26)$$

$$\{x: \sum_{i=0}^x b_{n,p}(i) \leq \alpha/2\} \cup \{x: \sum_{i=x}^n b_{n,p}(i) \leq \alpha/2\} \quad (\text{two-tailed test}) \quad (27)$$

The asymptotic one-sample test we study is the so-called score test. It may arise from inverting the confidence interval introduced by Wilson (Wilson, 1927) called the score interval. The alpha level critical region of the test is:

$$\{\hat{p}: \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq z_\alpha\} \quad (\text{left-tailed test}) \quad (28)$$

$$\{\hat{p}: \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_{1-\alpha}\} \quad (\text{right-tailed test}) \quad (29)$$

$$\{\hat{p}: \left| \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| \geq z_{1-\alpha/2}\} \quad (\text{right-tailed test}) \quad (30)$$

where \hat{p} is the observed proportion, p_0 is the null parameter value, n the sample size and z denote the percentile of the standard normal distribution corresponding to a certain alpha.

If the outcome is subject to misclassification with known Se and Sp (fixed parameter approach), the so-called Rogan and Gladen formula can be applied to calculate the true proportion from the observed one (Rogan & Gladen, 1978). The formula for this adjustment looks like

$$p_{adj} = (p_{obs} + Sp - 1) / (Se + Sp - 1), \quad (31)$$

where Se and Sp denote the sensitivity and specificity of the diagnostic test and p_{adj} and p_{obs} denote the adjusted and observed proportions. Reiczigel et al. (Reiczigel et al., 2010) showed that applying the same Rogan & Gladen formula to the endpoints of a confidence interval constructed for the sample proportion results in a valid confidence interval for the true proportion. Furthermore, the adjustment preserves exactness of the CI. These properties of the CIs have similar implications on testing.

The method to calculate the sample size in a scenario where the misclassification parameters are estimated from independent validation samples is based on inverting the confidence interval proposed by Lang and Reiczigel (Lang & Reiczigel, 2014). The detailed description of their method is included in section 2. of their article.

3.2.1.2 Two-sample tests

In a two-sample exposure-disease status association study, two different types of disease misclassifications can arise. Non-differential misclassification occurs when neither Se nor Sp for disease classification varies by exposure category (i.e. between the two groups we wish to compare) while differential misclassification occurs when misclassification of a disease status varies by exposure category (Q. Chen et al., 2013).

For the two-sample scenarios we used the method introduced by Hársfalvi and Singer (Hársfalvi & Singer, 2023) which is the combination of the method proposed by Zou and Donner (Zou & Donner, 2008) and one of the two methods from the ones proposed by Reiczigel et al. (Reiczigel et al., 2010) and Lang and Reiczigel (Lang & Reiczigel, 2014). With known sensitivity and specificity the Zou and Donner method was combined with the method proposed by Reiczigel et al. (Reiczigel et al., 2010). This is an asymptotic method that can be used in differential as well as non-differential misclassification scenarios. Let p_1 be the

population proportion in one and p_2 in the other group in a study. Using difference as effect measure, by inverting the confidence interval of Háršfalvi and Singer, we can test for: $H_0: p_1 - p_2 = 0$ against $H_a: p_1 - p_2 \neq 0$ (two-tailed test).

The first step of the Háršfalvi and Singer method is the construction of the confidence limits for the individual prevalence in each study group (p_1 and p_2) for which we use the previously presented Wilson method (Wilson, 1927), resulting in the initial lower and upper confidence limits for both proportions: (l_1, u_1) and (l_2, u_2) . Using the Rogan and Gladen formula the way proposed by Reiczigel we can obtain the adjusted confidence limits for the proportions in each group: (l'_1, u'_1) and (l'_2, u'_2) . If (l'_1, u'_1) and (l'_2, u'_2) are the confidence intervals of a certain level for the parameters p_1 and p_2 with estimates \hat{p}_1 and \hat{p}_2 being the Rogan-Gladen adjusted point estimates of p_1 and p_2 , then the confidence limits (LCL, UCL) for the risk difference $p_1 - p_2$ (having the same confidence level as the two independent confidence intervals) can be computed with the Zou and Donner formula:

$$LCL = \hat{p}_1 - \hat{p}_2 - \sqrt{(\hat{p}_1 - l'_1)^2 + (u'_2 - \hat{p}_2)^2} \quad (32)$$

$$UCL = \hat{p}_1 - \hat{p}_2 + \sqrt{(\hat{p}_2 - l'_2)^2 + (u'_1 - \hat{p}_1)^2} \quad (33)$$

The α level test for the risk difference can be obtained by inverting the above confidence interval.

For the two-sample test with estimated Se and Sp, the same approach was used. The Zou and Donner method was combined this time with the method proposed by Lang & Reiczigel (Lang & Reiczigel, 2014) instead of the one proposed by Reiczigel (Reiczigel et al., 2010).

3.2.2 The proposed method for handling potential drop-out patients

It is known that the power of the binomial test does not depend monotonically on sample size but displays a saw-tooth pattern (Chernick & Liu, 2002), thus, it may occur that for some n the power is above 80% but for a higher sample size ($> n$) it unexpectedly falls again under 80%. An example of this is shown in Figure 6 for the one-sample binomial test.

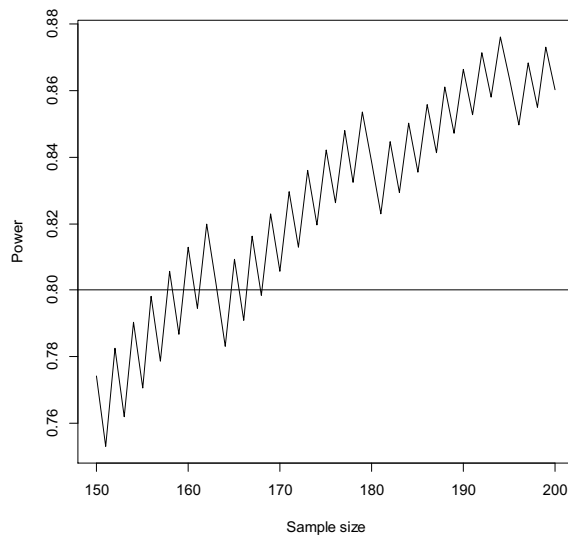


Figure 6. - Power of the exact binomial test is not a monotonic function of sample size
 $(p_0 = 0.5, p_a = 0.4, \text{alternative} = \text{"left-sided"}, Se = Sp = 1)$

Unfortunately, the actual sample size of a study, despite the hardest efforts, may differ from the planned one, and non-monotonicity of power invalidates the simplest method of handling this “to play safe, add 10% to the calculated sample size”, which works well for continuous outcomes. Even though the saw-tooth pattern of the power function is well known, it is easy to find clinical trials still using this simple but risky method of handling potential drop-out patients (clinicaltrials.gov, NCT01693614 and NCT02844582). The phenomena do not only occur for the one-sample case, the saw-tooth pattern is also present for a two-sample test.

To avoid this trap, some authors recommend choosing the smallest n so that for all $m \geq n$ the power is at least 80% (Chernick & Liu, 2002). However, it is a too strict requirement because if it can be ensured that the drop-out rate remains under a certain limit λ , say, under 5%, a smaller sample size than that may be sufficient. Therefore, we propose a sample size procedure that searches for the minimal sample size n so that even in case of some drop-out, not exceeding the specified proportion λ , the power never drops below the prescribed value. That is, for each sample size m from $(n - \lambda \cdot n)$ to n the power reaches the prescribed value, say 80%.

3.2.2 Study Settings

For the one-sample tests the selected null proportions p_0 in H_0 and assumed true proportions p_a are listed in Table 1 and we determined the necessary sample size n by exact power calculation. For each n we calculated the power so that we determined the alpha-level critical

region C of the test and calculated the probability of C assuming a binomial distribution with $p=p_a$. We calculated sample sizes for Se and Sp values 1, 0.99, 0.98, 0.95, 0.90.

As we suspected that the increase in the necessary sample size may differ for the two one-tailed tests (even for the two-tailed test depending on whether p_a is located left or right from p_0), we investigated each one separately. Thus, we set up two p_a for each p_0 : one left and the other right from p_0 (see Table 6). These were selected so that the sample size in case of no misclassification takes a few hundreds. We did not include p_0 values above 0.5 because results for $p_0 > 0.5$ are mirror-images of those for $p_0 < 0.5$. For example, power of test for $p_0=0.9$ with $p_a=0.96$, $Se=0.99$, and $Sp=0.95$ is same as that for $p_0=0.1$ with $p_a=0.04$, $Se =0.95$, and $Sp=0.99$.

Table 6. - Null and assumed true probabilities for the one-sample test. Left- and right-tailed tests were evaluated separately.

p_0	.01	.02	.03	.05	.10	.20	.30	.50
p_{aL}	.0005	.001	.003	.01	.04	.12	.20	.40
p_{aR}	.04	.07	.09	.12	.18	.32	.42	.62

Additionally, for the non-fixed misclassification parameter approach, we investigated how much the size of the validation samples of the Se and Sp impacts the study sample size, so we decided to use values of 25, 50, 100, 500 and 1000 for both n_{Se} and n_{Sp} .

For the analyses with fixed Se and Sp, our R function carries out the sample size calculation for five tests: the Clopper-Pearson exact test, the Wald-test, the Wilson's score test, the Agresti-Coull-test, and Blaker's exact test. It has an additional argument to specify the highest proportion of data loss λ (due to drop-out or other reasons), which still must not result in power less than the prescribed value. The function returns the minimal sample size n so that prescribed power is reached for each sample size from $(n - \lambda \cdot n)$ to n . In the present study we calculated sample sizes assuming a drop-out rate of $\lambda =0.15$, that is, with power remaining at least 80% up to 15% drop-out.

The function for the non-fixed parameter case can calculate the sample size with the option of handling the potential drop-outs the same way but are based on simulated power and currently only available for the two-sided test alternative.

For the fixed Se and Sp two-sample test we calculated the necessary sample sizes also for Se and Sp values of 1, 0.99, 0.98, 0.95 and 0.90 and used the same drop-out rate of $\lambda =0.15$,

assuming equal size of the two groups, and that drop-out may occur with equal chance in any group. We calculated power by simulation and considered only the two-sided alternative. The proportion pairs were also selected arbitrarily with the goal that the initial (without misclassification) sample sizes would take a few hundreds. The p_1 and p_2 pairs used are listed in Table 7.

The results can be reproduced using the functions available at the GitHub repository:

<https://github.com/Ragnar0ss/EffectofMisclassificationonSampleSize>.

Table 7. - The p_1 and p_2 pairs used in the evaluation of the two-sample test.

p_1	.01	.02	.03	.05	.10	.20	.30
p_2	.10	.12	.13	.15	.20	.35	.50

3.3 Results and Discussion

3.3.1 Study Results

As the first general observation we found that even small misclassification probabilities may result in considerable increase of sample size necessary to reach the prescribed power. For both the one and two-sample tests, the sample size increases the most if any of the two presumed probabilities (p_1 or p_2 for the two-sample and p_0 or p_a for the one-sample scenario) is extreme, that is, near the edges of the parameter space (to 0 or to 1). The results demonstrated that one extreme parameter is sufficient in case of misclassification to increase the required sample size magnificently.

The second general observation, applicable for all scenarios is that if the true prevalences are low (near 0) the variability of the estimated prevalence and in consequence the sample size is influenced more strongly by the Sp of the diagnostic test, whereas if the true prevalences are high (near 1) the sample size is influenced more strongly by Se. For medium prevalence (near 0.5) the influence of sensitivity and specificity are balanced.

For the one-sample exact test with fixed Se and Sp, Table 8 illustrates the dependence of sample size on p_0 , Se, and Sp in case of $Se = Sp$. Results for $p_0 > 0.5$ can be obtained by the symmetry argument as the scenarios for $p_0=p$, $Se=x$, $Sp=y$ and $p_0=1-p$, $Se=y$, $Sp=x$ are mirror images of each other.

Table 8. - Sample sizes for the one-sample exact test with fixed Se and Sp for all alternatives in scenarios with equal sensitivity and specificity

Se / Sp	p_0 (alternative = left-sided)								p_0 (alternative = right-sided)							
	0.01	0.02	0.03	0.05	0.1	0.2	0.3	0.5	0.01	0.02	0.03	0.05	0.1	0.2	0.3	0.5
1	352	176	185	146	168	180	153	199	197	132	133	132	152	109	130	136
0.99	1440	520	312	208	182	188	165	204	298	170	149	145	162	116	138	146
0.98	2376	762	436	269	206	202	176	217	409	213	175	166	176	122	140	149
0.95	5437	1526	840	456	289	249	197	242	732	323	259	223	218	139	162	169
Se / Sp	p_0 (alternative = two-sided with $p_a < p_0$)								p_0 (alternative = „two-sided” with $p_a > p_0$)							
	0.01	0.02	0.03	0.05	0.1	0.2	0.3	0.5	0.01	0.02	0.03	0.05	0.1	0.2	0.3	0.5
1	433	216	217	204	197	218	197	248	266	171	148	155	188	132	159	171
0.99	1830	623	386	284	224	238	208	260	379	203	176	176	202	138	164	182
0.98	2971	922	562	337	258	257	212	268	497	258	211	204	222	149	172	186
0.95	6846	1903	1051	568	353	305	250	303	905	402	317	270	271	177	199	212

Although the most dramatic effects of misclassification were observed when p_0 or p_a was near 0 or 1 (more than fourfold increase in sample size for $p_0 = 0.01$ with left-sided alternative and $Se = Sp = 99\%$), even in the best case, that is when $p_0 = 0.5$, the increase in necessary sample size was 22% with $Se = Sp = 95\%$, and 9% with $Se = Sp = 98\%$.

For the two-sided alternatives, results differ depending on whether the assumed true proportion p_a is smaller or greater than p_0 . Increase of necessary sample size is greater than that for the respective one-sided alternative illustrated by the extreme sample size increase of more than 6400 individuals (from an initial 433) in case of the two-sided alternative with $p_a = 0.0005$, $p_0 = 0.01$ and $Se = Sp = 95\%$.

Detailed calculation of the sample sizes for the most widely used asymptotic method, suggested by Wilson were also performed for the one-sample test. As expected, Wilson method resulted in a smaller sample size being less strict than the exact test by allowing a less rigorous control of the Type I error probability. Other tendencies are similar but slightly more extreme compared to those observed for the exact test (more than 4.5-times increase in sample size for $p_0 = 0.01$ with left-sided alternative and $Se = Sp = 99\%$), also in the best case, that is when $p_0 = 0.5$, the increase in necessary sample size was 26%. We also performed the same calculations for other asymptotic methods as well (Agresti-Coull and Wald) resulting in the same tendencies and similar amount of sample size increase.

For the two-sample test with fixed Se and Sp, apart from the fact that differential misclassification is possible, the observed changes are like the one-sample ones. Decreased Se and Sp increases the sample size required for the same power with the intensity of the increase depending on the position of the presumed parameters within the parameter space. For parameters closer to 0 the effect of the Sp is stronger, while for parameters closer to 1 the effect of the Se gets stronger. Additionally, the more the parameter values get closer to 0.5 the less the effect of both Se and Sp gets. In Table 9. we present a summary table for the two-sample test with fixed Se and Sp in scenarios with equal Se and Sp for both groups.

Table 9. - Sample size increase in % per group for the two-sample test with fixed Se and Sp in scenarios with equal Se and Sp for both groups For $Se_1=Se_2=Sp_1=Sp_2=1$ the required sample size is 110 as a baseline.

$p_1 = 0.01$ and $p_2 = 0.1$														
Se₁	Sp₁	Se₂	Sp₂	Sample Size	Se₁	Sp₁	Se₂	Sp₂	Sample Size	Se₁	Sp₁	Se₂	Sp₂	Sample Size
increase (%)					increase (%)					increase (%)				
1	1	0.99	0.99	5.5	0.99	0.99	1	1	17.3	0.99	0.99	0.99	0.99	24.5
1	1	0.98	0.98	17.3	0.98	0.98	1	1	34.5	0.98	0.98	0.98	0.98	49.1
1	1	0.95	0.95	44.5	0.95	0.95	1	1	78.2	0.95	0.95	0.95	0.95	125.5
1	1	0.9	0.9	116.4	0.9	0.9	1	1	176.4	0.9	0.9	0.9	0.9	299.1

The results so far represent the scenarios we assumed that both Se and Sp of the diagnostic test is known. Considering them as fixed values is seldom realistic as it ignores the uncertainty in the parameters, resulting in unreasonably low sample size estimates.

We found that the impact of Sp and Se on the sample size when these are estimated parameters is about the same as that with fixed parameters. If the true prevalence is low (near 0) the power of the test is influenced more strongly by Sp, whereas if the true prevalence is high (near 1) it is influenced more strongly by Se. Accordingly, the size of the validation sample for the parameter influencing the power more strongly has more relevance.

For example, in a testing scenario for the one-sample binomial test when the true parameter value (p_a) is closer to the lower end of the parameter space we know that Sp is the diagnostic parameter that will impact the power of the test more. As a consequence, to reduce the required sample size effectively, the size of the validation sample from which Sp is estimated should be increased while the size of the validation sample for the estimation of Se has only marginal effect.

Table 10 compares sample sizes in case of known vs estimated specificity for the one-sample test with $p_0 = 0.01$ and $p_a = 0.1$. For $n_{Sp} = 100$ and $Sp = 0.95$, the sample size needed is 674, nearly 16-fold of the fixed one, while for smaller validation studies ($n_{Sp} = 25$ or 50) the required sample size is completely unrealistic “> 10000”.

Table 10. - Sample sizes for the one-sample exact test with Se and Sp estimated from validation studies with different sample size, $p_0 = 0.01$ and $p_a = 0.1$ with $Se = 1$. The sample size needed for the scenarios with the same parameters but fixed Se and Sp are in the last row.

		Sp			
		1	0.99	0.98	0.95
n_{Sp}	25	> 10000	> 10000	> 10000	> 10000
	50	125	468	> 10000	> 10000
	100	55	84	122	674
	500	43	52	64	105
	1000	43	51	69	95
Sample size with fixed Se and Sp		43	51	67	94

What is also visible in Table 10 is that for higher Sp values ($Sp = 1, 0.99$ or 0.98), the sample size quickly gets closer to the fixed parameter values, even for moderate validation study sizes ($n_{Sp} = 100$), while with lower Sp values it only coming close to the fix scenario for large validation study sizes ($n_{Sp} = 500-1000$).

The same findings were identified with the two-sample approach with estimated Se and Sp. As there are more sources of variability by allowing four parameters to be estimated from validation samples (Se_1, Se_2, Sp_1 and Sp_2), the sample size can increase more easily, but the same rules introduced earlier apply to determine which parameter influences the study the most. As an example, if the expected true proportions for the two groups are $p_1 = 0.13$ and $p_2 = 0.03$, the parameters having the most impact on the sample size will be Sp_2 and n_{Sp_2} , as p_2 is the parameter closer to the edge of the parameter space with Sp being the parameter more impactful at the lower edge.

We performed all calculations for different drop-out rates, alpha error rates and power settings as well. Apart from the obvious impact of these on the sample size, the effects of misclassification were similar in all different scenarios, namely that the more alpha error rate is acceptable the less the sample size is while for higher power or with higher expected drop-out rate, a higher sample size is required.

Our results showed that ignoring even small misclassification probabilities may result in considerable power loss. As most clinical research studies do not take misclassification into account in the design phase, the conventional sample size calculation will result in a study conducted with less power than preferred. Thus, potential misclassification must be considered in sample size calculation for studies with binomial endpoints even if sensitivity and specificity are high (98-99%), necessary sample sizes may be multiples of those without misclassification.

3.3.1 Case Studies

We generated a case study for both testing scenarios to illustrate how the design can change based on the assumptions one has about the potential misclassifications.

Example 1. – Two-sample test

In a randomized, double blind, active controlled study we wish to evaluate the safety and efficacy of an antiviral agent in comparison to the standard-of-care therapy. The study is designed to use the same rapid test in both groups to determine the ratio of patients with a positive hepatitis B surface antigen as the primary endpoint of the study. In the design phase, the expected true proportions of each group are $p_1 = 0.4$ and $p_2 = 0.25$ with the expected drop-out rate being between 5 and 15%.

By ignoring potential misclassification (assuming $Se = Sp = 1$ in both groups) and drop-out, the necessary sample size for each group is 152 patients with $\alpha = 0.05$ to reach 80% power.

Now let us assume that we plan to use the HBsAg Hepatitis B Antigen Test by RightSign in both groups to detect the disease status and consider the diagnostic parameters of the rapid test as fixed values ($Se = 0.998$ and $Sp = 0.996$ from the website of the manufacturer). If we do not take account of the drop-out, the necessary sample size to reach the same 80% power is slightly larger, 155 subjects / group.

If we continue to ignore the potential drop-outs and consider both Se and Sp as random variables and consider the sample sizes of the validation studies the manufacturer used with $n_{Sp} = 1612$ and $n_{Se} = 425$ (User Manual Number RP5303400 with Effective date: 2019-07-26), the sample size for the main study increases to 159 / group. The large size of the validation studies controls the sample size increase.

If we want to protect against the drop-out and consider both Se and Sp as random variables, the sample size increases further based on the expected drop-out rate. With 5% the sample size increases to 166 / group while with 15% the study size increases to 185 / group.

Example 2. – One-sample test

Our goal is to evaluate the free from infection status from Mycobacterium Tuberculosis in a Member State or a zone by investigating whether the incidence rate of the Mycobacterium Tuberculosis infection in the establishments during a year is lower than 2% ($p_0 = 0.02$). For this, we use the exact binomial test. The expected infection rate is $p_a = 0.002$ with a drop-out rate between 20% and 40%.

Without misclassification and by ignoring the potential drop-outs, the sample size needed to reach 80% power in a study with $\alpha = 0.05$ is 278 in total.

Using a rapid test on the market, the LIODetect TB-ST Tuberculosis Rapid Test with diagnostic parameters as fixed, Se = 0.6535 and Sp = 0.9659 while still ignoring potential drop-outs, the required sample size increases to an extreme 2536 in total.

Using the latest approach but taking drop-outs into account the total sample size increases to 3347 with 20% expected drop-outs and 4462 with 40% expected drop-outs.

With estimated Se and Sp parameters and moderate validation study sizes used by the manufacturer $n_{Sp} = 323$ and $n_{Se} = 153$ (Instructions for Use Rev. 3.0 / 181019), the sample size gets quite unrealistic, with > 50000 even when potential drop-outs are ignored.

4 Applying the promising zone method for sample size re-estimation in clinical trials when the binomial endpoint is based on a diagnostic test

4.1 Background

Adaptive clinical trial designs allow for preplanned mid-trial modifications to one or more elements of the study design based on the data that is already collected in the study (Bauer et al., 2016). The “promising zone” approach based on conditional power is one of these adaptive sample size re-estimation designs, first proposed by Chen et al. in 2004 (Y. J. Chen et al., 2004), broadened by Gao et al. (Gao et al., 2008) and made more “accessible to practitioners” by Mehta and Pocock in 2011 (C. R. Mehta & Pocock, 2011). The conditional power (CPo) is defined as the conditional probability that the null hypothesis will be rejected at the end of the study given that a particular result ($Z_1=z_1$) is observed at the interim analysis. With this design, if the sample size is re-calculated only when the CPo is in the promising zone, then the re-assessment will not increase the overall type I error rate.

From all different adaptive designs submitted for regulatory review, the most popular have been the ones with sample size re-estimation (SSRE), according to a survey of scientific advice letters from the European Medicines Agency (Elsässer et al., 2014). Edwards et al (Edwards et al., 2020) performed a systematic review on the application of the “promising zone” design and concluded that it is being implemented in clinical trial practice for a wide range of situations, being especially appealing to researchers due to its simplicity.

When designing a traditional fixed-sized clinical trial with a binomial endpoint prone to misclassification, the sample size evaluation is usually based on the assumptions for the true theoretical response rates, disregarding the effect of misclassification. At the final analysis, observed results are adjusted for sensitivity and specificity. For more flexible designs (either sequential or adaptive) the same adjustments should also be made during the interim analyses (He et al., 2021).

The promising zone depends on the information fraction on which the sample size re-assessment is based, on the maximum of the permitted increase in sample size, and on the targeted conditional power. In the design phase it is assumed that this CPo is based on the true response rates, disregarding the effect of misclassification. However, the conditional power calculated at the interim is based on the observed data, the so-called “trend” CPo according to Lachin’s (Lachin, 2005) terminology. To be consistent with the initial power calculations, the observed CPo should always be adjusted for the imperfection of the diagnostic test (sensitivity and specificity) used for the observation of the interim results.

4.2 Methods

For a simple presentation of the promising zone method first described by Chen (Y. J. Chen et al., 2004) and popularized by Mehta (C. R. Mehta & Pocock, 2011), let's assume a two-stage clinical trial setting where patients enter the trial one at a time and are randomized to the experimental or the control group. Subject outcomes are assumed to be independent, identically distributed random variables and we also assume normality with equal variances. The null hypothesis to be tested at an alpha level is $H_0: \delta = 0$, against the one-sided alternative $H_a: \delta > 0$, where δ is the treatment difference between the means of the control and experimental group.

The following notations are introduced: n_1 is the sample size planned for the first stage (prior to the preplanned interim analysis), \tilde{n}_2 is the originally planned incremental sample size for the second stage ($n_2 = n_1 + \tilde{n}_2$, the total sample size for the trial), z_1 is the test statistic for stage one, z_α is the $(1-\alpha)$ -quantile of the standard normal distribution, and ϕ is the cumulative distribution function of the standard normal distribution.

Increasing the sample size based on the results of an unblinded interim analysis may inflate the type I error rate and to control it at the nominal level, appropriate adjustments are needed (Y. J. Chen et al., 2004). One option to maintain the nominal level of the type I error rate is to use a weighted Wald-type test statistic, as suggested by Cui (Cui et al., 1999). The specific weights for the two stages are $\frac{n_1}{n_2}$ for the first and $\frac{\tilde{n}_2}{n_2}$ for the second stage, often called as the CHW statistic. If these weights are used to sum up the two Z values from the two stages, then the type I error rate is preserved. If the weight for the second stage is modified based on the new, re-calculated sample size \tilde{n}_2^* to $\frac{\tilde{n}_2^*}{n_2}$, then the boundary for the weighted sum (and therefore the significance level) would also have to be modified to preserve the type I error rate, and the modification is given by the following formula:

$$b(z_1, \tilde{n}_2^*) = (\tilde{n}_2^*)^{-0.5} \left[\sqrt{\frac{\tilde{n}_2^*}{\tilde{n}_2} (z_\alpha \sqrt{n_2} - z_1 \sqrt{n_1})} + z_1 \sqrt{n_1} \right] \quad (34)$$

The conditional power at the interim is calculated based on the Z value obtained in the first stage, $Z_1 = z_1$, by the following formula:

$$CPo(z_1, \tilde{n}_2) = 1 - \phi \left(\frac{z_\alpha \sqrt{n_1 + \tilde{n}_2} - z_1 \sqrt{n_1}}{\sqrt{\tilde{n}_2}} - \frac{z_1 \sqrt{\tilde{n}_2}}{\sqrt{n_1}} \right) \quad (35)$$

The simple and elegant idea was to define a promising zone where no correction is needed for the critical value of the Z statistic because in this region $b(z_1, \tilde{n}_2^*) \leq z_\alpha$ holds. In this region the boundary calculated by (34) does not exceed the originally planned z_α , therefore the type

I error rate is preserved if the original z_α is used as a boundary instead of $b(z_1, \tilde{n}_2^*)$, allowing to use conventional test statistics instead of the weighted ones.

The general principle is to define three different zones for the conditional power (unfavorable, promising, and favorable) in advance and assign specific boundaries to each. At the interim analysis the conditional power is calculated based on the observed data ($CP_{\delta_1}(z_1, \tilde{n}_2)$) and the three zones are defined as:

- *unfavorable zone* ($CP_{\delta_1}(z_1, \tilde{n}_2) < CP_{min}$), continue with the originally calculated sample size as the results are disappointing
- *favorable zone* ($CP_{\delta_1}(z_1, \tilde{n}_2) > 1 - \beta$), also continue with the originally calculated sample size as there is no need to adaptively increase it
- *promising zone* ($CP_{min} \leq CP_{\delta_1}(z_1, \tilde{n}_2) \leq 1 - \beta$), increase the sample size within the pre-specified limits to recover the targeted power to the planned level

Although more versatile design options are available allowing researchers to either stop the trial early for efficacy or futility (C. Mehta et al., 2022), in our review the original settings of the method will be used.

Now let's assume a binomial outcome based on a diagnostic test with known Se and Sp causing potential misclassification, and a single-arm clinical trial with null hypothesis $H_0: p = p_0$ to be tested against the one-sided alternative $H_a: p > p_0$, where p is the true response rate of the treatment. As discussed earlier, to get an unbiased estimate of the treatment effect and an accurate conditional power at the interim analysis, Z_1 should be calculated with proper adjustments around the settings of the diagnostic test (Se and Sp):

$$Z_{1,adj} = \frac{\hat{p}_{1,adj} - p_0}{\sqrt{\frac{\hat{p}_{1,adj}(1 - \hat{p}_{1,adj})}{n_1}}} \quad (36)$$

Applying to the observed treatment response rate the adjustment first proposed by Rogan and Gladen (Rogan & Gladen, 1978) when the Se and Sp are considered to be known constants results in:

$$\hat{p}_{1,adj} = \frac{\hat{p}_{1,obs} + Sp - 1}{Se + Sp - 1} \quad (37)$$

with this adjustment inserted to (36), the sample size for the second stage can be re-calculated as:

$$\tilde{n}_2^* = \frac{n_1}{Z_{1,adj}^2} \left[\frac{z_\alpha \sqrt{n_2} - Z_{1,adj} \sqrt{n_1}}{\sqrt{n_2 - n_1}} + z_\beta \right]^2 \quad (38)$$

4.3 Results and Discussion

First, we illustrate the method described above by applying it to a confirmatory medical device clinical trial and later formulate some general thoughts on the effect of misclassification to the promising zone adaptive design. Some pieces of the example are constructed but some are from an actual trial, however, for confidentiality reasons the trial details had to be omitted.

4.3.1 Example - a medical device trial for spinal disorder

A medical device manufacturer wanted to test his transforaminal lumbar interbody fusion implants for surgical lumbar spine stabilization. Fusion at month 24 was the primary outcome variable, and the statistic of interest was the cumulative fusion rate at month 24. This rate was intended to be compared to the state-of-the-art industry standard, which was 90%. A right-sided one sample binomial test was planned to be applied for this purpose ($\alpha=0.025$, $\beta=0.2$). The originally planned trial size, $n_2=78$ was obtained based on the assumption that the new device has a 98% cumulative fusion rate at month 24 and aimed to attain an 80% power to detect superiority compared to the state-of-the-art standard (90%) at a one-sided 0.025 level, applying a Wald-type statistic.

As the manufacturer was not sure about the true performance of the implants, and lower cumulative fusion rates would also be considered clinically meaningful, other options (92%, 94% and 96%) were investigated as well.

Due to the above-mentioned uncertainties the manufacturer decided to consider the adaptive approach. The plan was to start a trial targeting the originally planned total number of subjects as 78 ($n_2 = 78$), perform an interim analysis after observing $n_1 = 50$ completers and have the possibility to increase the sample size if the interim results are promising. The maximal sample size was selected to be $n_{max} = 300$ and the sponsor had no intention to stop early (either due to futility or efficacy). Under these conditions the pre-determined zones were the following:

- *Favorable zone*: If $CP \geq 0.8$, continue until the enrollment of $n_2 = 78$ patients.
- *Promising zone*: If $0.365 \leq CP \leq 0.8$, increase the sample size to $n_2^* = \min(n_2', 300)$, where n_2' is the re-calculated sample size to achieve $CP=0.8$.
- *Unfavorable zone*: If $CP \leq 0.365$, continue until the enrollment of $n_2 = 78$ patients.

An additional uncertainty was around the possible misclassification of the fusion status. Potential misclassification may increase the uncertainty of the conclusion and reduce the power of the test, so usually the necessary sample size needed for the same power is higher than it would be without misclassification.

After investigating the expected sample sizes for each adaptive scenario with the previously identified parameters (fusion rate, Se and Sp), there is no important difference between the scenarios where the conditional power was calculated with the adjusted response rate compared to the ones without adjustment. The real advantage of using the approach with adjustment can be seen from Figure 7 where the probability of each interim outcome zone is shown with two different true fusion rates.

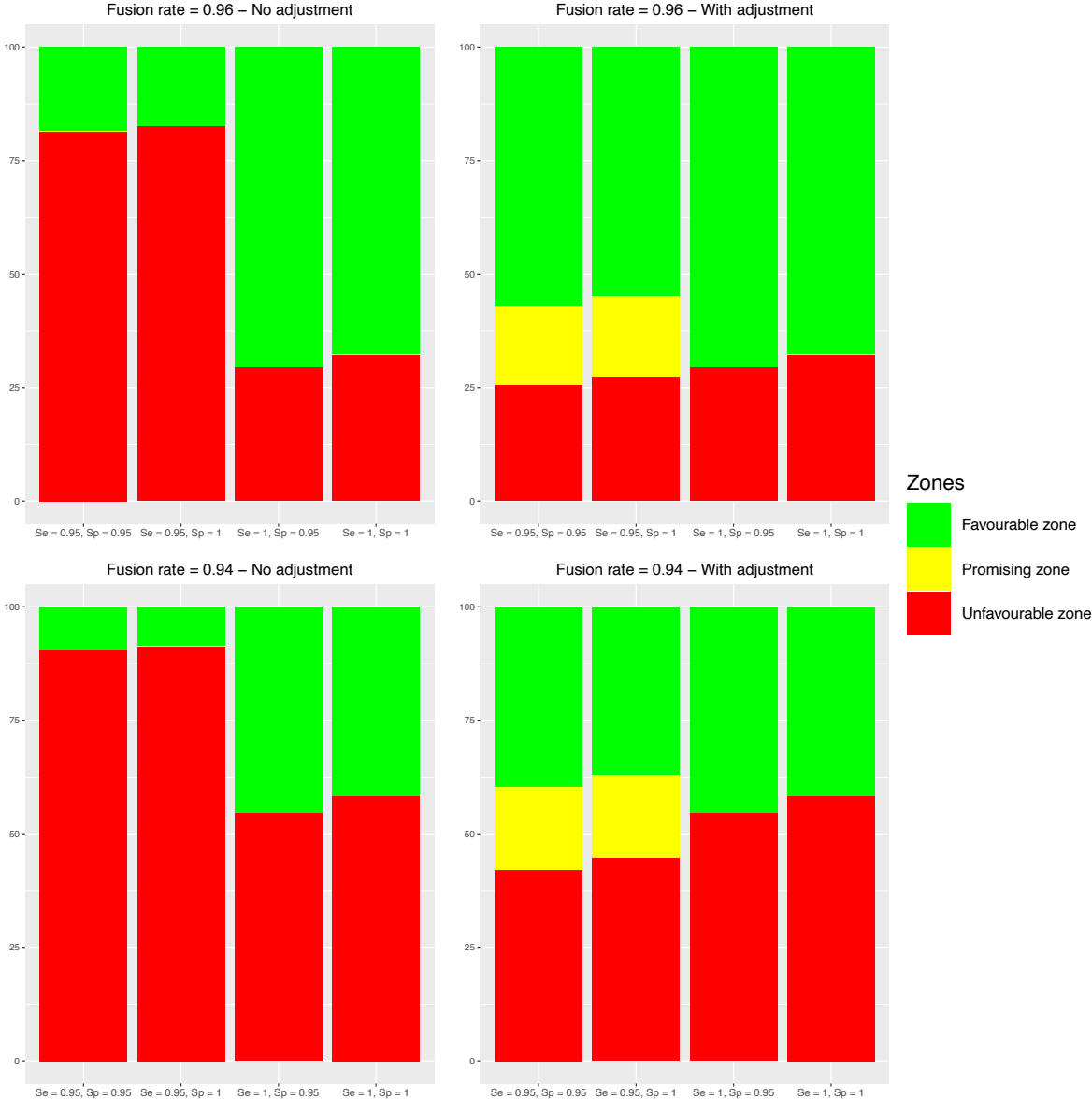


Figure 7. - Probability of each interim outcome zone with and without adjustment for two different true fusion rates (all results are based on 1.000.000 simulated trials)

With the use of the adjustment the probability of a successful trial increases significantly. It is also clear from Figure 7 that the effect is impacted more by the sensitivity as expected and is the most extreme with rates near the edge of the parameter space.

4.3.2 General thoughts on the impact of misclassification

An important difference between continuous endpoints and the binomial endpoints prone to misclassification is that the range of observed response rates (and as a consequence, the range of Z_1 values) is limited by the sensitivity and specificity of the diagnostic test (Y. J. Chen et al., 2004; C. R. Mehta & Pocock, 2011). From the formula

$$p_{obs} = p_{adj}Se + (1 - p_{adj})(1 - Sp), \quad (39)$$

it can be deduced that the observed response rate ranges from $1-Sp$ (when $p_{adj} = 0$) to Se (when $p_{adj} = 1$).

To illustrate how dramatically an imperfect diagnostic test can alter the observed test statistics during an interim analysis, in Figure 8, Z_1 values unadjusted, and adjusted for sensitivity and specificity are presented, for various null hypotheses and different observed response rates for an interim analysis performed at $n_1=50$.

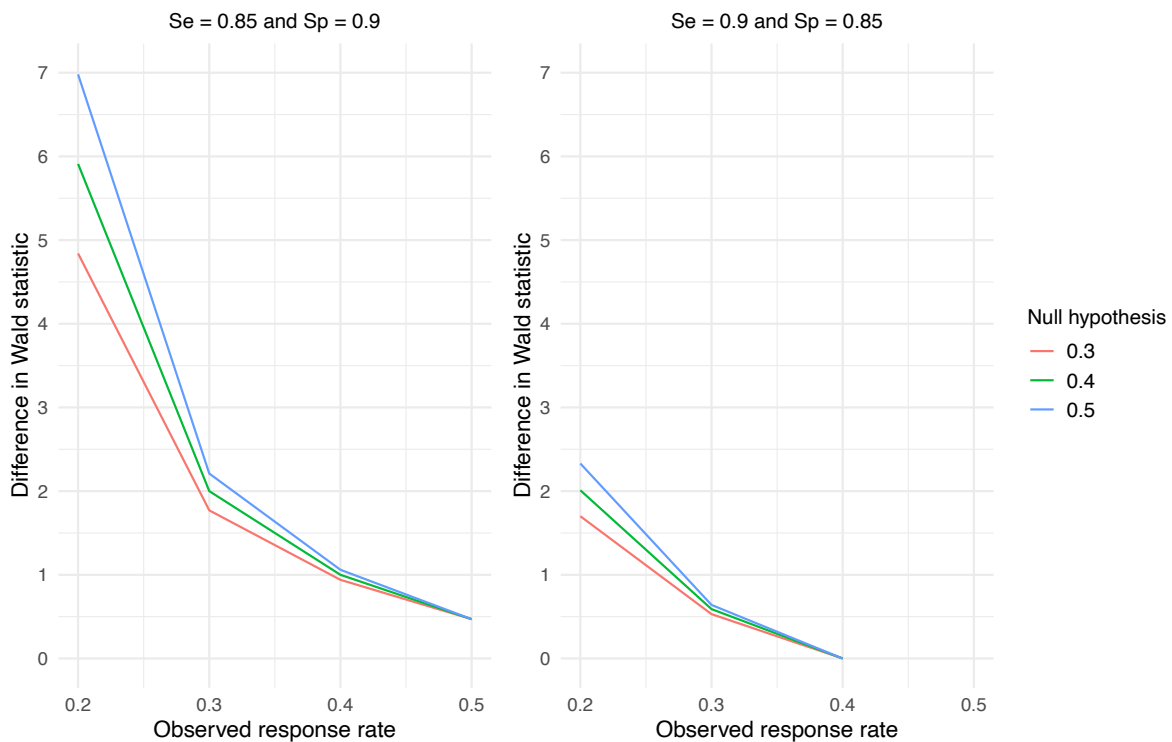


Figure 8. - Adjusted and unadjusted Z_1 -values for different null hypotheses and different response rates, for $n_1=50$, $Se=0.9$ and $Sp=0.85$

As shown in Figure 8, discrepancies between the observed Z_1 scores can be quite large, therefore differences between the corresponding conditional power values are expected to be also significant.

The range of differences between adjusted and unadjusted conditional power was computed over all the possible CPo values (between 0.1 and 0.9 by steps of 0.001) for various scenarios defined by the null hypothesis, the information fraction where the interim analysis was performed, and the maximum sample size allowed after re-calculation for stage 2. All other parameters were left unchanged (sensitivity of the diagnostic test is 0.9, whereas the specificity is 0.85, the originally planned sample size $n_2=200$ and one-sided type I error rate is 0.025 and type II error rate to be 0.1). Some of the calculations are presented below in Table 11. Differences less than 10% are illustrated with pale green background, yellow shading represents a difference of at least 10%, orange a difference of at least 20% while red background represents a difference greater than 30%.

Table 11. - The range of differences between the adjusted and unadjusted CPo (%) values for $n_2=200$, $Se=0.9$ and $Sp=0.85$

Null hypothesis	Maximum incr. allowed (n_2^*/n_2)	Information fraction at interim (n_1/n_2)	Minimum diff. between adjusted and unadjusted CP (%)	Maximum diff. between adjusted and unadjusted CP (%)
$p_0 = 0.4$	1.5	0.25	-46.7	-9.92
		0.50	-55.1	-9.98
		0.75	-71.1	-10.0
	2.0	0.25	-46.7	-9.92
		0.50	-55.1	-9.98
		0.75	-71.1	-10.0
	2.5	0.25	-46.7	-9.92
		0.50	-55.1	-9.98
		0.75	-71.1	-10.0
$p_0 = 0.5$	1.5	0.25	-15.5	1.19
		0.50	-17.8	-0.17
		0.75	-23.0	-3.74
	2.0	0.25	-15.5	1.19
		0.50	-17.8	-0.17
		0.75	-23.0	-3.74
	2.5	0.25	-15.5	1.19
		0.50	-17.8	-0.17
		0.75	-23.0	-3.74
$p_0 = 0.6$	1.5	0.25	7.18	29.04
		0.50	9.75	34.24
		0.75	9.93	45.64
	2.0	0.25	7.18	29.04

		0.50	9.75	34.24
		0.75	9.93	45.64
	2.5	0.25	7.18	29.04
		0.50	9.75	34.24
		0.75	9.93	45.64

As reflected by Table 11, the smallest maximum differences were found for the null hypothesis of 0.5 and the biggest for $p_0=0.6$. The maximum increase allowed did not influence the difference between the adjusted and unadjusted CPo values, whereas differences showed an increasing trend with when the interim analysis occurred. Closer the interim analysis was to the originally planned sample size the bigger the difference between the adjusted and unadjusted CPo values got. For the same scenarios the promising zone intervals were also calculated but no significant differences were found between their adjusted and unadjusted version.

Ignoring diagnostic uncertainty may have a considerable effect on the power of a clinical trial. Adaptive study designs such as the „promising zone” approach are an emerging trend for developers to increase the likelihood of study success, but when a diagnostic test is used in the observation of the outcome the results can be immensely flawed. As illustrated, an imperfect diagnostic test influences the observed Wald statistic and also the conditional power calculated during an interim look on the data.

An adjustment easy to implement, yielding an unbiased estimate of the treatment effect and an accurate conditional power at the interim analysis is presented in this paper. Our proposal is to always use the adjustment when the sensitivity and specificity of a diagnostic test applied for the classification of the outcome is less than 1.

5 Logistic regression with covariate-dependent probability of misclassification

5.1 Background

Logistic regression is one of the most important methods of categorical data analysis. It is used to model the probability of an event as a function of some predictors in a variety of applications including biomedical studies, social science research and business applications (Agresti, 2012).

In traditional analysis, it is assumed that the variables in the model are measured perfectly and do not contain any measurement error, which rarely occurs in practice. Both the outcome variable and / or some predictors may be subject to errors. For the outcome, being a qualitative variable, the measurement error manifests itself as misclassification. Misclassification may be due to imperfect diagnostic procedures, recall bias, or people's reluctance to speak honestly about sensitive issues (Bollinger & David, 1997; Egleston et al., 2011; Gorber et al., 2009; Tourangeau & Yan, 2007).

Many authors have investigated the logistic regression model while it is exposed to some kind of measurement error. Ignoring these measurement errors completely can lead to serious biases, erroneous regression coefficients and standard error estimates that jeopardize the validity of the statistical inference in the study (Carroll et al., 1995, 1995; Hausman et al., 1998; Y. Liu et al., 2013; Luan et al., 2005).

Others (Magder & Hughes, 1997) studied the effects of measuring imperfectly the dependent variable in logistic regression and proposed a method based on the expected-maximization (EM) algorithm to account for this when misclassification rates are known. Their method produced unbiased estimates of the odds ratios with a greater variance than estimates ignoring the test imperfections.

In (Hausman et al., 1998) the authors came up with a modified ML estimator similarly for a misclassified dependent variable, combining the maximum rank correlation estimator of Han with isotonic regression (Han, 1987). Gustafson et al. also worked on the problem when classification probabilities are considered as known and proposed a Bayesian method to incorporate uncertainty in the model (Gustafson, 2003). For other closed form corrections that can be used for the logistic regression model, see (Duffy et al., 2004) and (Greenland, 2008).

Neuhaus presented an expression that quantifies the loss of information due to misclassification. He showed that ignoring response misclassification leads to attenuated covariate effects only when the errors are independent of the covariates, while if misclassification probabilities depend on covariates, the errors can lead to extreme bias (Neuhaus, 1999). His expression can also be used as a sample size inflation factor when designing studies where one will most probably have misclassified responses. Later Neuhaus touched on the subject of covariate-dependent misclassification and concluded that if the misclassification probabilities depend on covariates, the probability of response will depend on these same covariates through a linear predictor and the misclassification probabilities (Neuhaus, 2002).

Many authors proposed methods to handle misclassification using internal or external validation samples (Cheng & Hsueh, 1999; Küchenhoff et al., 2006; Lyles et al., 2011; Spiegelman et al., 1997, 2000). Lyles et al. demonstrated how misclassification rates can depend on the values of subject-specific covariates and illustrated the importance of accounting for this dependence (Lyles et al., 2011). They also stated that models that include sensitivity and specificity without validation data may not be identifiable or may be misleading in the case of model misspecification, but there are situations where validation data are rather difficult to obtain, especially in social science applications. Spiegelman et al. maximized likelihood from both the main study and a validation study to obtain maximum likelihood estimates for the parameters of the underlying logistic regression model, the measurement error model and reclassification models simultaneously. In this work, he proposed the optimal size for such a validation study (Spiegelman et al., 2000).

Davidov et al. studied the effect of outcome-dependent misclassification, in which the misclassification of a binary covariate in a logistic regression model depend on the observed outcome (Davidov et al., 2003). They examined covariate-dependent misclassification of exposures and of outcomes.

Liu and Zhang proposed a method that allows joint estimation of regression coefficients and misclassification rates without validation data (H. Liu & Zhang, 2017). The Liu-Zhang model corresponds to the four-parameter logistic model used in pharmacology, toxicology and psychology (Barton & Lord, 1981; Waller & Feuerstahler, 2017).

In the item response theory, a branch of psychometrics, the Liu-Zhang model may describe a situation when correctly answering an item of a knowledge test is, besides dependence on the knowledge level, subject to both false positivity due to knowledge-free guessing, and false

negativity due to inattention. In such context, Martinková and Hladká proposed a kind of Liu-Zhang model generalization as a tool for detection of differential item functioning (Martinková & Hladká, 2023). Considered is a pair of proband groups, denoted as reference and focal group. The probability of correctly answering an item is modelled by Liu-Zhang model, but with different parameters in each group. Thus, within the overall model common for both groups, the probabilities of false positive and false negative result and other model elements are covariate-dependent, where the covariate is a binary variable coding the group membership. Further generalization to a more complex group structure is possible.

Our aim was to extend the logistic model to situations where the dependence of misclassification rates on some covariates can be described by logistic regression. It is well documented that sensitivity and specificity of a diagnostic test may depend on several factors (Braden & Caspary, 2001; Coughlin et al., 1992; Toft et al., 2005; von Euler-Chelpin et al., 2019). We propose a logistic regression model that allows for misclassification of the outcome, in which either sensitivity or specificity may depend on some covariates while the other one is assumed to be constant. The proposed model is a generalization of the Liu-Zhang model.

Our model can be applied when the presence of a trait (seropositivity, disease, etc.) depends on some covariates and its probability of detection (if present) depends on other covariates. It may also have applications in social sciences. For some sensitive survey questions, respondents are reluctant to answer honestly, and the degree of honesty may depend on certain covariates. In such cases our model allows the simultaneous estimation of the true proportion conditional on the covariates and the degree of response bias conditional on its covariates.

5.2 Methods

5.2.1 Description of the model

For the formal description of the model let Y_{true} and Y_{obs} denote the true and the observed response, respectively, and let Se denote the sensitivity or detection probability, and Sp denote the specificity. In terms of misclassification rates, $1 - Se$ and $1 - Sp$ are the misclassification rates given $Y_{true} = 1$ and $Y_{true} = 0$ respectively.

Assume that the probability of $Y_{true} = 1$ depends on some predictors X_i while sensitivity depends on some other predictors Z_j , each according to a logistic model. Specificity is assumed to be a constant. A more formal model (1) description is as follows:

$$\begin{aligned}
Y_{true} &\sim \text{Bernoulli}(p), \\
\text{logit}(p) &= \beta_0 + \sum_{i=1}^r \beta_i X_i, \\
Se &= P(Y_{obs} = 1 | Y_{true} = 1), \\
\text{logit}(Se) &= \gamma_0 + \sum_{j=1}^q \gamma_j Z_j, \\
Sp &= P(Y_{obs} = 0 | Y_{true} = 0) = \text{constant}, \\
Y_{obs} &\sim \text{Bernoulli}(pSe + (1-p)(1-Sp)).
\end{aligned} \tag{Model 1}$$

Notation may be simplified from X_1 to X when $r = 1$, and similarly from Z_1 to Z when $q = 1$. Changing the codes of the outcome $Y_{obs.new} = 1 - Y_{obs}$ turns the model to another model in which specificity is covariate-dependent and sensitivity is constant.

5.2.2 Parameter estimation

Model 1 can be fitted to observed data by maximum likelihood (ML). The parameters of the model are $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_r)$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_q)$, and Sp . Their likelihood for a single observation, depending on the observed outcome y_{obs} , is as follows.

If $y_{obs} = 1$, then for $\mathbf{x} = (x_0, x_1, \dots, x_r)$, $\mathbf{z} = (z_0, z_1, \dots, z_q)$:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, Sp, \mathbf{x}, \mathbf{z}, y_{obs}) = p(\boldsymbol{\beta}, \mathbf{x})Se(\boldsymbol{\gamma}, \mathbf{z}) + (1 - p(\boldsymbol{\beta}, \mathbf{x}))(1 - Sp), \tag{40}$$

and if $y_{obs} = 0$, then:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, Sp, \mathbf{x}, \mathbf{z}, y_{obs}) = p(\boldsymbol{\beta}, \mathbf{x})(1 - Se(\boldsymbol{\gamma}, \mathbf{z})) + (1 - p(\boldsymbol{\beta}, \mathbf{x}))Sp, \tag{41}$$

where

$$p(\boldsymbol{\beta}, \mathbf{x}) = \frac{\exp(\beta_0 + \sum_{j=1}^r \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^r \beta_j x_j)}, \tag{42}$$

and

$$Se(\boldsymbol{\gamma}, \mathbf{z}) = \frac{\exp(\gamma_0 + \sum_{j=1}^q \gamma_j z_j)}{1 + \exp(\gamma_0 + \sum_{j=1}^q \gamma_j z_j)}. \quad (43)$$

For a sample of mutually independent observations, the likelihood function is naturally the product of the likelihoods of individual cases.

The model enables ML estimating the true probability of the studied feature conditional on the covariates X_i as well as the sensitivity (probability of detection) conditional on its covariates Z_i via numerical nonlinear optimization. Standard errors of estimates can be estimated by jackknife.

5.2.3 Sub-model testing

The principal tool for testing hypotheses about parameters of model 1 is the LRT. To compare model (1) with other models that are not its sub-models, we use Akaike's Information Criterion (AIC) (Akaike, 1974).

Though the crucial question related to model 1 is, as a rule, whether Y_{true} depends on X_1, \dots, X_r , i. e. whether $\beta_1 = \dots = \beta_r = 0$, other hypotheses and various sub-models of model (1) may also be of interest.

The matter is a bit simpler in case of single covariates for both Y_{true} and Se (which is the case in power study described in the next section, as well as in the real-data examples than for more complex settings, so that we split exposition into two branches.

5.2.3.1. Case $r = q = 1$

The full model (1) has 5 parameters, $\beta_0, \beta_1, \gamma_0, \gamma_1$ and Sp . This five-parameter model will be called as $M5$.

The question arises whether all parameters are necessary to describe the data, or a simpler model may fit (almost) equally well, and serve also as a base for testing hypothesis $\beta_1 = 0$.

It is relevant comparing $M5$ with two four-parameter sub-models. One of them is the Liu-Zhang model (to be called as LZ), which assumes constant sensitivity, that is, $\gamma_1 = 0$. The other four-

parameter model to be compared with $M5$ is that with the constraint $Sp = 1$. We call this latter model $M4$.

If both models fit significantly worse than $M5$, then the only possibility is to test the dependence of Y_{true} on X within $M5$, that is, compare $M5$ with its submodel in which $\beta_1 = 0$.

If LZ does not fit worse than $M5$, then it should be compared with its sub-model in which $Sp = 1$. If $M4$ does not fit worse than $M5$, then it should be compared with its sub-model in which $\gamma_1 = 0$. Both constraints, $Sp = 1$ in LZ and $\gamma_1 = 0$ in $M4$ result in the same three-parameter model that we denote by $M3$.

Finally, if $M3$ is not significantly worse than any of $M4$ and LZ , then it is to be compared with the basic logistic model without misclassification, that is, in which both Se and Sp are equal to 1. We denote this two-parameter model by $M2$. When $M2$ is not significantly worse than $M3$, then it proves to be sufficient model for the data, and distinguishing Y_{true} from Y_{obs} is found unnecessary.

The above model comparisons are summarized in Figure 9.

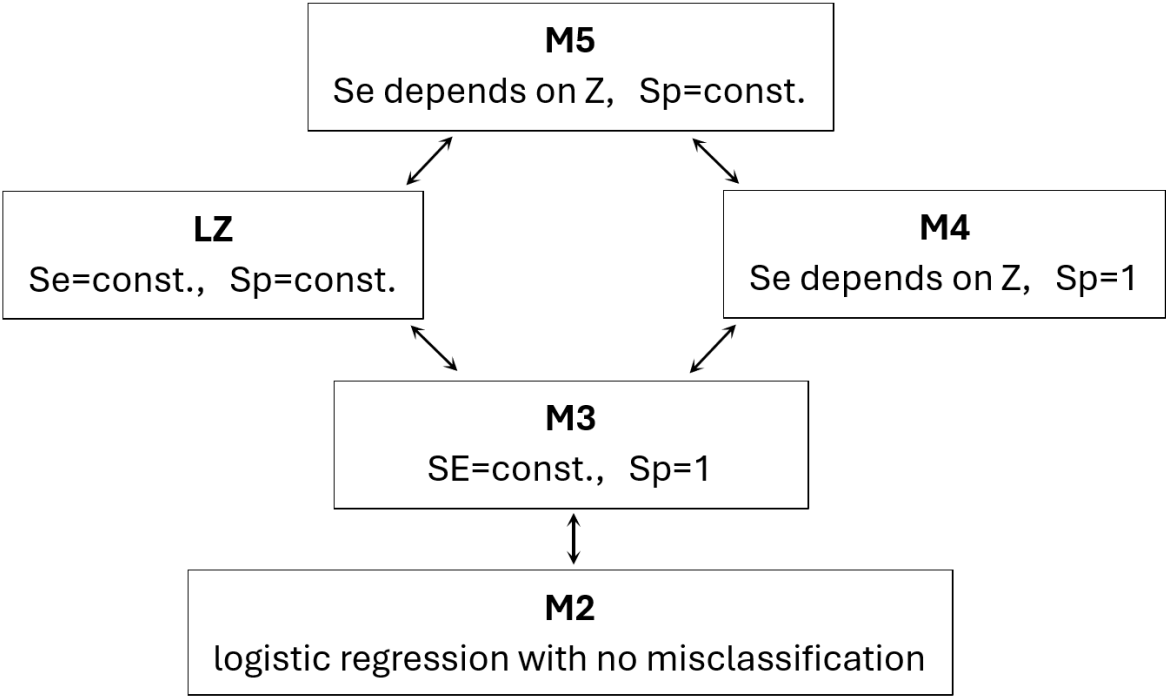


Figure 9. – Diagram of the recommended model comparisons

5.2.3.2. Case $r > 1$ and / or $q > 1$

Analogous model hierarchy as shown in Figure 9 takes place in the more complex setting, and analogous steps should be performed as those in Section 5.2.3.1. Only instead of scalar-valued variables $X = X_1, Z = Z_1$ and scalar parameters β_1, γ_1 we have now vector-valued ones $\mathbf{X} = (X_1, \dots, X_r), \mathbf{Z} = (Z_1, \dots, Z_q)$ and vectors of parameters $\boldsymbol{\beta}^* = (\beta_1, \dots, \beta_r), \boldsymbol{\gamma}^* = (\gamma_1, \dots, \gamma_q)$. Obviously, where testing of $\beta_1 = 0$, or $\gamma_1 = 0$ is required in 5.2.3.1., now $\boldsymbol{\beta}^* = (0, \dots, 0)$, or $\boldsymbol{\gamma}^* = (0, \dots, 0)$, respectively, should be tested via LR test.

Notations $M5, M4$ and other makes sense even here (and may be kept), if we take each of $\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*$ as one vector parameter.

The only thing relevant here that has no counterpart in 5.2.3.1. is the search for nonzero components of $\boldsymbol{\beta}^*$ or $\boldsymbol{\gamma}^*$ in case that hypothesis $\boldsymbol{\beta}^* = (0, \dots, 0)$, or $\boldsymbol{\gamma}^* = (0, \dots, 0)$, respectively, is rejected. This can be done in the same way as for the usual logistic regression model.

5.2.4 Identifiability issues

Let us consider a model slightly more general Model 1, namely:

$$\begin{aligned}
 Y_{true} &\sim \text{Bernoulli}(p), \\
 \text{logit}(p) &= \beta_0 + \sum_{i=1}^r \beta_i X_i, \\
 Se &= P(Y_{obs} = 1 | Y_{true} = 1), \\
 \text{logit}(Se) &= \gamma_0 + \sum_{i=1}^r \gamma_i X_i, \\
 Sp &= P(Y_{obs} = 0 | Y_{true} = 0) = \text{constant}, \\
 \text{logit}(Sp) &= \delta_0 + \sum_{i=1}^r \delta_i X_i \\
 Y_{obs} &\sim \text{Bernoulli}(pSe + (1-p)(1-Sp)).
 \end{aligned}
 \tag{Model 2}$$

The difference between models (1) and (2) is that under model (2), both Se and Sp may be covariate-dependent, and moreover Se and Sp may according to model (2) depend on the same variables as Y_{true} . Model (1) can be obtained from (2) through constraints on parameters, when (i) all δ_i except for $i = 0$ are set to zeroes and, moreover, (ii) for each $i \geq 1$, either γ_i or β_i is constrained to zero.

Let $f(x)$ denote the logistic function of real x , i.e.

$$f(x) = \exp(x) / (1 + \exp(x)), \quad (44)$$

and further for $\mathbf{x} = (x_1, \dots, x_r)$ and $\boldsymbol{\xi} = (\xi_0, \xi_1, \dots, \xi_r)$ let

$$g(\mathbf{x}, \boldsymbol{\xi}) = f(\xi_0 + \sum_{i=1}^r \xi_i x_i). \quad (45)$$

When $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_r)$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_r)$ and $\boldsymbol{\delta} = (\delta_0, \delta_1, \dots, \delta_r)$, then in case that $X_i = x_i$ for all $i \geq 1$, under model (2) Se, Sp and probability $p = P_{\boldsymbol{\beta}}(Y_{true} = 1)$ can be expressed as:

$$Se = g(\mathbf{x}, \boldsymbol{\gamma}), Sp = g(\mathbf{x}, \boldsymbol{\delta}), p = g(\mathbf{x}, \boldsymbol{\beta}), \quad (46)$$

and probability that $Y_{obs} = 1$ as

$$P_{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}}(Y_{obs} = 1) = h(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = g(\mathbf{x}, \boldsymbol{\gamma})g(\mathbf{x}, \boldsymbol{\beta}) + (1 - g(\mathbf{x}, \boldsymbol{\delta}))(1 - g(\mathbf{x}, \boldsymbol{\beta})). \quad (47)$$

Since $f(-x) = 1 - f(x)$, it is easy to see that

$$h(\mathbf{x}, -\boldsymbol{\beta}, -\boldsymbol{\gamma}, -\boldsymbol{\delta}) = h(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}). \quad (48)$$

This means that model (2), if unconstrained, is not identifiable, since the distribution of Y_{obs} for parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ is exactly the same as for $\boldsymbol{\beta}' = -\boldsymbol{\beta}$, $\boldsymbol{\gamma}' = -\boldsymbol{\delta}$, and $\boldsymbol{\delta}' = -\boldsymbol{\gamma}$.

The issue is well known in the context of the Liu-Zhang model (Hausman et al., 1998; H. Liu & Zhang, 2017) a sub-model of model (2) with $\gamma_i = \delta_i = 0$ for all $i \geq 1$. In that case, there is a standard remedy, based on the fact that for $Se = g(\mathbf{x}, \boldsymbol{\gamma})$, $Sp = g(\mathbf{x}, \boldsymbol{\delta})$, $Se' = g(\mathbf{x}, \boldsymbol{\gamma}')$ and $Sp' = g(\mathbf{x}, \boldsymbol{\delta}')$, where $\boldsymbol{\gamma}' = -\boldsymbol{\delta}$, and $-\boldsymbol{\gamma} = \boldsymbol{\delta}'$, identity

$$Se + Sp + Se' + Sp' = 2, \quad (49)$$

Holds, so that only one of (constant) sums $Se + Sp$ and $Se' + Sp'$ may exceed 1. Then, only such parameter values are considered acceptable (which appears reasonable) that make the sum of Se and Sp greater than 1. (When $Se + Sp = Se' + Sp' = 1$, both variants yield the same degenerate model.) Identity holds not only for the Liu-Zhang model but for model (2) in general, however, the sum of Se and Sp is not constant under model (2), and there is no universal rule that could make it uniformly high enough, regardless of variables, X_1, \dots, X_r . One of possible

solutions could consist in requiring that the sum of Se and Sp exceeds 1 for a fixed combination $\mathbf{x} = (x_1, \dots, x_r)$ of values of variables X_i , i.e.

$$g(\mathbf{x}, \boldsymbol{\gamma}) + g(\mathbf{x}, \boldsymbol{\delta}) = \gamma_0 + \delta_0 + \sum_{i=1}^r (\beta_i + \delta_i)x_i > 1. \quad (50)$$

When, for an instance, applied to zero levels of all variables X_i , the condition reduces to $\gamma_0 + \delta_0 > 1$. Nevertheless, the choice of vector \mathbf{x} or, more generally, of solution that would make model (2) identifiable (i. e. a rule that for each pair $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}), (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\delta}') = (-\boldsymbol{\beta}, -\boldsymbol{\gamma}, -\boldsymbol{\delta})$ of parameter space elements decides which one to take as acceptable, and which to discard) may rather depend on the specific field where the model is applied.

The problem becomes easier when, which is the case in model (1), a constant Sp is assumed. Transformation $\boldsymbol{\beta}' = -\boldsymbol{\beta}, \boldsymbol{\gamma}' = -\boldsymbol{\delta}, \boldsymbol{\delta}' = -\boldsymbol{\gamma}$ yields a model with constant Sp only when Se appears constant as well, so that the model reduces to the Liu-Zhang one. In such a case, the only remaining concern, as regards identifiability, is to force the sum of Se and Sp to exceed 1.

However, yet another identifiability issue arises when Sp , either by assumption, or as a result of numerical search, equals 1. The probability that $Y_{obs} = 1$ then reduces to

$$h(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = g(\mathbf{x}, \boldsymbol{\gamma})g(\mathbf{x}, \boldsymbol{\beta}). \quad (51)$$

So, the same distribution of Y_{obs} as for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is obtained for $\boldsymbol{\beta}' = \boldsymbol{\gamma}$ and $\boldsymbol{\beta} = \boldsymbol{\gamma}'$. In other words, it is impossible to determine which of the parameter vectors $\boldsymbol{\beta}, \boldsymbol{\gamma}$ 'belongs to' Y_{true} , and which to Se . Thus, if no constraints are applied to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the model is unidentifiable. The problem vanishes if, as in model (1) Se and Y_{true} are assumed to depend on different variables, that is for each $i \geq 1$, one of the parameters β_i, γ_i is zero.

Let us take a look at the question whether, provided that Sp uniformly equals 1, it is possible that at least some of variables X_1, \dots, X_r influence both Y_{true} and Se , and the model is still identifiable. Let us classify variables X_i that influence Y_{true} or Se into sets A, B and AB , depending on which of parameters β_i, γ_i are, and which are not constrained to zero. In A , β_i is, unlike γ_i , while in B , by contrast, γ_i is and β_i is not. Finally, for X_i in AB neither of β_i, γ_i is constrained to zero. By assumption, AB and at least one of A, B are nonempty. It may seem that transformation $\boldsymbol{\beta}' = \boldsymbol{\gamma}$ and $\boldsymbol{\beta} = \boldsymbol{\gamma}'$ cannot preserve the constraints on β_i for X_i from A , and on γ_i for X_i from B . However, this holds only as long as some of these parameters take on

nonzero values. When $\gamma_i = 0$ for all X_i from A and $\beta_i = 0$ for all X_i from B (and it would be very unusual not to allow that, i. e. not to consider such possibility), $\beta' = \gamma$ and $\beta = \gamma'$ satisfy all constraints defining the classification into A and B . Then, since parameters β_i generally do not equal γ_i for X_i from AB , also (β', γ') differs from (β, γ) , and the model is not identifiable.

The conclusion is that if Sp is or may be the unite constant then, for the sake of model identifiability, the lists of variables on which Se and Y_{true} depend should, as is the case in model (1), be disjunctive.

However, this conclusion can be somewhat weakened in some cases if we abandon the purist view that the model must be identifiable in every case and look at the matter more pragmatically. In case that the unit specificity is considered just as a theoretical possibility in the framework of a more general model, and is extremely weakly supported by the data, the given kind of identifiability issue may be of only limited importance from the practical point of view. Since the problem concerns just a small subset of parameter space and the procedure of likelihood maximization is likely to converge to a solution laying elsewhere. Similarly, if at unit specificity the data strongly contradict zero values of parameters β_i, γ_i for all X_i from groups denoted above as A and B , the result of ML search, as well as its interpretation, might be unambiguous even if some covariates X_i influence both Y_{true} and Se .

Nevertheless, the possibility of such a weakening depends on specific conditions, and if we want a safe and general rule, the previous strict conclusion applies.

5.3 Results and Discussion

5.3.1 Power assessment

5.3.1.1. Structure of experiments

We made simulation experiments to explore the statistical power of the method to detect in the setting with $r = q = 1$ the dependence of outcome on X and sensitivity on Z . The power of logistic regression depends on the true parameters, the distribution of predictors, and the sample size (Væth & Skovlund, 2004). For any sample size, the highest power is obtained if data cover the whole range from small to high outcome probabilities, but this condition is not met in most observational studies. Which part of the logistic curve is covered by the data depends on the response rate in the range of the linear predictor. In case of our model, the

situation is more complicated because we have two logistic curves that are linked together. Therefore, we tried to explore by simulation how the power depends on the range of the predictors X and Z , Sp , and response rates.

We made experiments in which:

- data covered nearly the whole probability range for both logistic curves (1 experiment)
- data covered the lower, middle, and upper probability range for each curve (9 experiments – 3 * 3)
- data covered the lowest, lower, middle, higher, highest probability range for each curve (25 experiments – 5 * 5)

Table 12 shows the definition of these ranges in terms of probability ranges as well as ranges of the linear predictor for an increasing logistic curve. Simulation was carried out with increasing logistic curves, that is, with data from the model introduced as model (1) with $\beta_1 = \gamma_1 = 1$ and $\beta_0 = \gamma_0 = 0$. Each experiment was repeated for 5 values of Sp (0.7, 0.8, 0.9, 0.999, 1) and for 5 different sample sizes: 500, 1000, 2000, 5000, and 10000, resulting in altogether $35 * 5 * 5 = 875$ simulation runs. In each run, alpha was set to 0.05 and power was evaluated from 500 replications, that is, 500 random datasets from the above model.

Values of X and Z were generated from normal distributions with mean value equal to the midpoint of the range, and standard deviation a quarter of the range width, resulting in about 95% of the values lying in the ranges given in Table 12. The results from these 700 simulation runs together with the functions are available at the GitHub repository: <https://github.com/Ragnar0ss/CovariateDependentLogisticRegression>.

A short summary is given in the following section.

Table 12. - Response rate ranges and ranges of linear predictor in italics used in the simulation experiments.

Experiments					
1-range	(0.01, 0.99) (-4.6, 4.6)				
3-range	(0.01, 0.5) (-4.6, 0)	(0.25, 0.75) (-1.1, 1.1)	(0.5, 0.99) (0, 4.6)		
5-range	(0.01, 0.3) (-4.6, -0.85)	(0.2, 0.5) (-1.39, 0)	(0.35, 0.65) (-0.62, 0.62)	(0.5, 0.8) (0, 1.39)	(0.7, 0.99) (0.85, 4.6)

5.3.1.2. Simulation results

In the following description of the simulation results, since we used increasing logistic curves, high X range corresponds to high probability range of Y , and high Z range corresponds to high sensitivity range.

1-range experiment

In the experiment covering the whole probability range for both logistic curves the power of detecting dependence of both Y on X and Se on Z exceeded 99% already for $N = 500$. Power of detecting that Sp is less than 100% was 89%, 98%, and 99% when Sp was 0.9, 0.8, and 0.7, respectively, and it improved with increasing sample size. For $N = 1000$ power reached 99% in all cases.

3-range experiments

Power of detecting dependence of Y on X was highest for high Z range and high Sp . Even for $N = 500$, it reached 99%. However, in case of low Z combined with lower Sp , it was as low as 30%. Results were similar for $N = 1000$ but here the power also reached 99% for middle Z range and also for lower Sp if both of these did not occur at the same time. For larger samples, power increased but for low Z with low Sp it remained rather low (37%, 41%, and 46% for $N = 2000, 5000,$ and 10000 , respectively).

Power of detecting dependence of Se on Z was highest for high X range. Even for $N = 500$ it reached 99%, independent of the Z range and Sp . The lowest power (less than 30%) occurred in case of low X range combined with low Z range and low Sp . For higher samples, power gradually increased. It reached 99% for middle X range ($N = 1000$ and 2000) and also for low X range ($N = 5000$ and 10000). Low X with low Z and low Sp remained the worst case with power estimates of 30%, 33%, 53%, and 83% for $N = 1000, 2000, 5000,$ and 10000 , respectively.

Power of detecting that Sp is less than 100% proved to be much lower than those above. For $N = 500$, its mean as well as median was 25%, and it exceeded 50% only when range of X was low and range of Z was high. Results for $N = 1000$ were similar. Mean and median of the power were 30% and 23%, respectively. Power exceeded 60% only if range of X was low and range of Z was high. For larger samples power showed some increase but even for $N = 10000$ its mean and median was about 60%. Highest power (for low X and high Z) reached 90% for $N = 2000$ and 99% for $N = 5000$ and 10000 .

5-range experiments

Power of detecting dependence of Y on X is highest for the lowest X range, highest Z range, and high Sp . For $N = 500$ the power may reach 99%. However, in case of low Z and low Sp , it may fall below 25%. For $N = 2000$, the results were the same, but the effect of the X range was smaller. For $N = 5000$ and 10000 , results were same except that the dependence of power on the X range disappeared. Here too, in case of low Z and low Sp , power was rather low, about 25 to 30% ($N = 5000$) and 25 to 40% ($N = 10000$). For the visual presentation of some results Table 13 displays the dependence of power on the outcome probability and Se ranges and Sp for a sample size of $N = 1000$.

Table 13. - Power of detecting dependence of Y on X for various outcome probability and Se ranges and Sp in the 5-range experiments for a sample size of $N=1000$. Each number is estimated from 500 replications.

Sp = 1		Se probability range				
		lowest	low	mid	high	highest
Y prob. range	lowest	0.676	0.982	1	1	1
	low	0.342	0.698	0.880	0.948	0.994
	mid	0.290	0.602	0.770	0.892	0.988
	high	0.274	0.566	0.714	0.890	0.994
	highest	0.242	0.550	0.732	0.884	0.996
Sp = 0.9		Se probability range				
		lowest	low	mid	high	highest
Y prob. range	lowest	0.280	0.524	0.786	0.936	0.996
	low	0.234	0.456	0.694	0.866	0.982
	mid	0.208	0.418	0.638	0.826	0.952
	high	0.192	0.416	0.582	0.822	0.974
	highest	0.286	0.374	0.622	0.794	0.992
Sp = 0.8		Se probability range				
		lowest	low	mid	high	highest
Y prob. range	lowest	0.310	0.340	0.536	0.752	0.952
	low	0.272	0.334	0.508	0.738	0.902
	mid	0.218	0.314	0.470	0.686	0.922
	high	0.226	0.256	0.468	0.698	0.954
	highest	0.294	0.268	0.460	0.716	0.978
Sp = 0.7		Se probability range				
		lowest	low	mid	high	highest
Y prob. range	lowest	0.432	0.274	0.332	0.566	0.858
	low	0.354	0.248	0.402	0.558	0.850
	mid	0.330	0.274	0.362	0.548	0.868
	high	0.358	0.240	0.320	0.546	0.906
	highest	0.364	0.242	0.374	0.552	0.950

Power of detecting dependence of sensitivity on Z is highest for the highest X range, lowest Z range, and high Sp . For $N = 500$ it may reach 99%. However, in case of the lowest X range, independent of the Z range and Sp , it may fall below 30%. For $N = 2000$, the same results were obtained. For $N = 5000$ and 10000 , power exceeded 95% in case of high Sp for all ranges of X except the lowest one. No dependence on Z was experienced. For the lowest X range combined with Sp less than 0.95, the power was on average as low as about 40%.

Power of detecting that Sp is less than 100% proved to be much lower than those above. For $N = 500$, its mean, median and maximum was 19%, 18%, and 33%, respectively, and it exceeded 30% only in 3 of the 75 experiments (here we omitted the experiments with Sp 0.999). Power showed no marked dependence on the X and Z ranges, nor on Sp . For $N = 2000$, power improved only for the lowest X range combined with the highest Z range, where it reached 55%, otherwise it was rather low (mean: 16%, median: 15%, upper quartile 20%). For $N = 5000$ and 10000 , it reached 90% for the lowest X range combined with the highest Z range but for other combinations it remained low, its mean as well as its upper quartile was 22% for $N = 5000$, and its mean was 27%, upper quartile 28% for $N = 10000$.

5.3.2 Applications

Age-dependence of alcohol consumption

For this example, we use the data from the CDC 2021 Youth Risk Behavior Survey (<https://www.cdc.gov/healthyyouth/data/yrbs/data.html>) to investigate age-dependence of alcohol consumption (Ada, 2020). Age distribution of the respondents is shown in Table 14. Due to the small frequencies in the first two categories, in the analysis we used the range from 14 to 18.

Table 14. - Age distribution of the respondents.

Age (years)	≤12	13	14	15	16	17	≥18
Frequency	39	62	3403	4427	4276	3904	1023

Alcohol consumption (A/c) was recoded from **Question 43** of the questionnaire:

43. During the past 30 days, what is the largest number of alcoholic drinks you had in a row, that is, within a couple of hours?

- A. I did not drink alcohol during the past 30 days (**Alc=0**),
- B. 1 or 2 drinks, C. 3 drinks, D. 4 drinks, ... (**Alc=1**).

Sensitivity (probability of honest reporting of drinking) was assumed to depend on the openly admitted alcohol consumption in the social environment of the responder, based on **Question 9**:

9. *During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol?*

- A. 0 times, B. 1 time, C. 2 or 3 times, D. 4 or 5 times, E. 6 or more times
- (Rode.w.drinking.driver = 1...5)**

The formal description of the model we fitted is

$$\begin{aligned} \text{logit}(Alc) &= \beta_0 + \beta_1 Age, \\ \text{logit}(Se) &= \gamma_0 + \gamma_1 Rode.w.drinking.driver. \end{aligned}$$

Including specificity, the model *M5* has 5 parameters to estimate. After omission of data with missing values, there remained 12163 responders. Fitting the model *M5* to data resulted in $\beta_0 = -7.60$, $\beta_1 = 0.47$, $\gamma_0 = -3.98$, $\gamma_1 = 3.41$, $Sp = 1$ (deviance = 11480.0, AIC = 11490.0). The corresponding logistic curves are shown in Figure 10. Parameter estimates as well as deviance of model *M4* are the same but its AIC is smaller by 2 because *M4* has one less parameter.

Specificity of 1 means that no one reports falsely alcohol consumption if it does not occur. At the same time, the dependence of sensitivity on *Rode.w.drinking.driver* shows that about 60% of those who did not report riding in a car with a drinking driver might drink but deny it. Those who admit riding sometimes with drinking drivers, do not hide alcohol consumption.

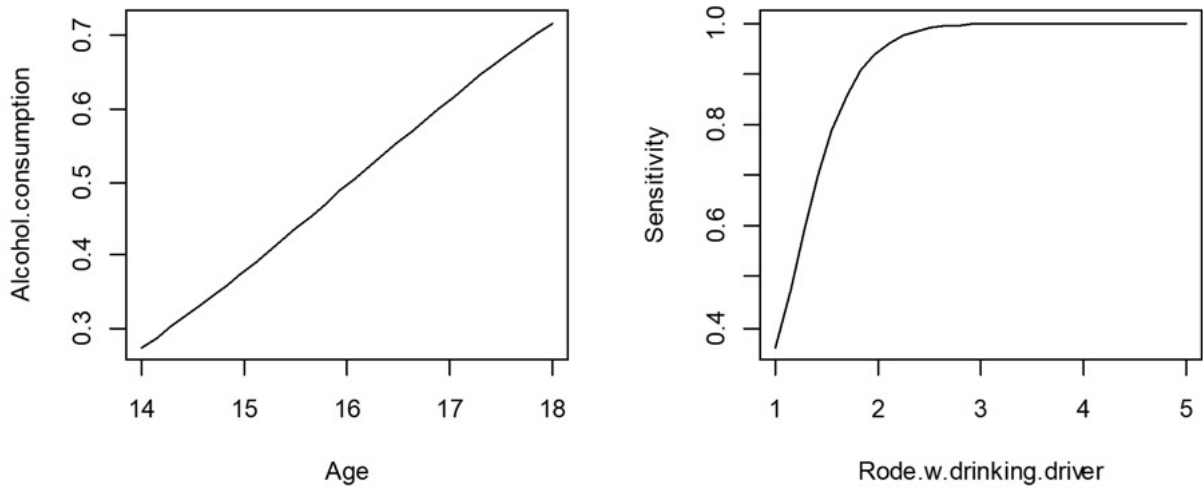


Figure 10. - Estimated dependence of the prevalence of alcohol consumption on age, and dependence of Se (honest reporting of alcohol consumption) on how often the respondent admittedly rode a car with a drinking driver. The estimated Sp is 1, which means that those respondents who do not consume alcohol are assumed to report their status accurately.

The submodel with $\gamma_1 = 0$ (constant sensitivity, that is, the Liu-Zhang model) resulted in $\beta_0 = -6.61$, $\beta_1 = 0.36$, $Se = 0.73$, $Sp = 1$ (deviance = 12111.7, AIC = 12117.7). This model fitted significantly worse than $M5$ ($p < 0.0001$). Assuming constant sensitivity instead of covariate-dependent one underestimates the prevalence of alcohol consumption, especially at younger ages. The corresponding logistic curve is shown in the left panel of Figure 11.

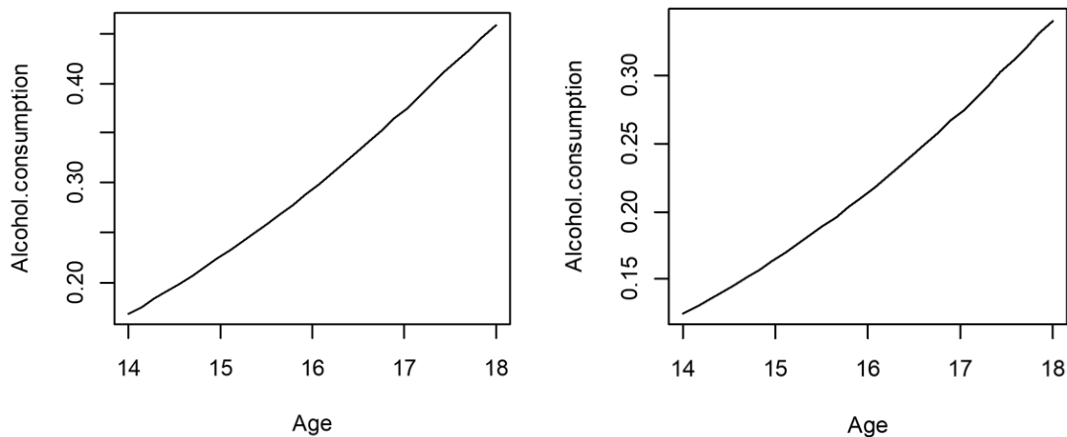


Figure 11. - Estimated age dependence of the prevalence of alcohol consumption assuming constant Se , that is, applying the Liu-Zhang model (left panel). The estimates of Se and Sp are $Se = 0.73$ and $Sp = 1$. Estimated age-dependence of the prevalence of alcohol consumption assuming no misclassification (right panel).

The submodel with $Se = Sp = 1$ resulted in $\beta_0 = -6.45$ and $\beta_1 = 0.32$ (deviance = 12111.8, AIC = 12115.8). Thus, the fit of this model and that of the Liu-Zhang model to data do not differ significantly ($p = 0.7198$), which means that the Liu-Zhang model does not deliver strong evidence for the occurrence of misclassification. The logistic curve with no misclassification is shown in the right panel of Figure 11.

Some may think that it is not sensitivity that is affected by the social environment but alcohol consumption itself. To test this, we fitted the logistic regression model with two explanatory variables: Age and Rode.w.drinking.driver. This model had AIC = 11697.4, which is 207.4 more than the AIC of our model with covariate-dependent sensitivity; that is, the data do not support this view.

Investigating voting habits in Europe

In the second application we fit a model in which sensitivity is assumed to be constant and specificity depends on a covariate. Here we use the following questions of the 2017 European Values Study (Ada, 2020).

- **Question 40** (in the role of X_1) - How important is it for you to live in a country that is governed democratically? On a scale from 1 to 10 where 1 means it is "not at all important" and 10 means "absolutely important" what position would you choose?
- **Question 44** (in the role of Z_1) - Please tell me for each of the following whether you think it can always be justified, never be justified, or something in between, using this card (from 1 to 10, where 1 means "never" and 10 means "always"). We chose the item "Someone accepting a bribe in the course of their duties".
- **Question 48** (in the role of Y) When elections take place, do you vote always, usually or never? Please tell me separately for each of the following levels (local, national, European). We chose the item "National level", and dichotomized the answers by combining categories "usually" and "never".

We presumed that among those who find democracy important, the proportion of those who always vote is higher. Further, we hypothesized that people who usually but do not always vote may answer "always" instead of "usually"; and the probability of this may correlate with how important they feel that one should always do the right thing. Thus, the more someone denounces bribery, the more likely they answer "always", forgetting about those occasions when they did not vote.

In this example it is easier to describe the model using misclassification probabilities instead of sensitivity and specificity, which represent probabilities of correct classification.

- $Y = P(\text{someone always votes})$ is an increasing function of the Question 40 value, that is, the higher someone values democracy, the higher is this probability.
- $P(\text{someone answers "not always"} \mid \text{always votes}) = (1 - \text{sensitivity})$ is constant.
- $P(\text{someone answers "always"} \mid \text{does not always vote}) = (1 - \text{specificity})$ is a decreasing function of the Question 44 value (the more one denounces bribery, the higher is this probability).

We fitted the model $M5$ to data of all 36 countries included in the European Values Study. Dependence of Y on Question 40 value was significant for 32 countries (exceptions: Belarus, Bosnia and Hercegovina, Montenegro, and Russia). Comparing $M5$ to $M4$, we found that Se did not differ significantly from 1 in 29 countries, and in 17 of them the estimated Se was equal to 1. This means that the estimated probability that someone answers "not always" when in fact always votes is in 17 countries equal to 0 and in further 12 countries does not differ significantly from 0.

Dependence of Sp on Question 44 answer was significant for 24 countries, which means that for these countries $M5$ fitted significantly better than the Liu-Zhang model that assumes constant specificity. For 27 countries, our model also fitted significantly better than the logistic regression model with two explanatory variables, Question 40 and Question 44, which suggest that Question 44 might affect the probability of false reporting rather than influence Y directly.

In the following, the results for the Czech Republic are presented in detail. The parameters of model $M5$ were $\beta_0 = 8.78$, $\beta_1 = -0.82$, $\gamma_0 = 0.023$, $\gamma_1 = 0.177$, and $Se = 1$. Deviance and AIC were 2208.11 and 2218.11, respectively. Both dependencies (that of "always voting" on the reported importance of democracy and that of falsely reporting "always voting" on the reported denunciation of bribery) were significant ($p < 0.0001$). Not surprisingly, as $M5$ estimated $Se = 1$, fitting $M4$ yielded the same parameter estimates and deviance as $M5$.

The LZ model, with estimated $Se = 0.637$ and $Sp = 0.593$, fitted significantly worse according to the likelihood ratio test ($p < 0.0001$). Deviance and AIC were 2227.84 and 2235.83, respectively. Standard logistic regression with two predictors (importance of democracy and denunciation of bribery), assuming a direct effect of denunciation of bribery on voting behaviour, yielded an AIC of 2233.17, 15.06 higher than the AIC of $M5$.

The results did not change substantially if we used “local level” as well as “European level” in Question 48. Also, we got similar results with other items in Question 44, for example “Cheating on tax if you have the chance”, or “Avoiding a fare on public transport”. A summary is given in Table 15.

Table 15. - Each of the 15 items in Question 44 was tested on the Czech data, whether the misclassification probability $P(\text{reports always} \mid \text{not always votes})$ depends on the answer to it. The dependence was significant for 5 items, highlighted in bold. All these 5 items are related to „comme il faut” behavior, and 4 of them are associated with minor misconduct against the state.

Item	p-value
Claiming state benefints which you are not entitled to	<0.0001
Cheating on tax if you have the chance	<0.0001
Taking the drugs marijuana or hashish	0.1059
Someone accepting a bribe in the course of their duties	<0.0001
Homosexuality	0.5192
Abortion	0.5679
Divorce	0.4687
Euthanasia (terminating the life of the incurably sick)	0.5648
Suicide	0.2394
Having casual sex	0.0318
Avoiding a fare on public transport	0.0030
Prostitution	0.0873
Artificial insemination or in-vitro fertilization	0.6833
Political violence	0.1657
Death penalty	0.3491

Figure 12 displays the logistic curves resulting from our model $M5$ and those from the Liu-Zhang model. Our model predicts that the proportion of those who always vote varies from 0 to 35%, depending on how important they feel democracy is. The estimate of Se is 1, that is, the model predicts that all those who always vote report this correctly. Those who do not always vote, however, may falsely report „always voting”. The probability of this varies between 15% and 45%, depending on their moral attitude.

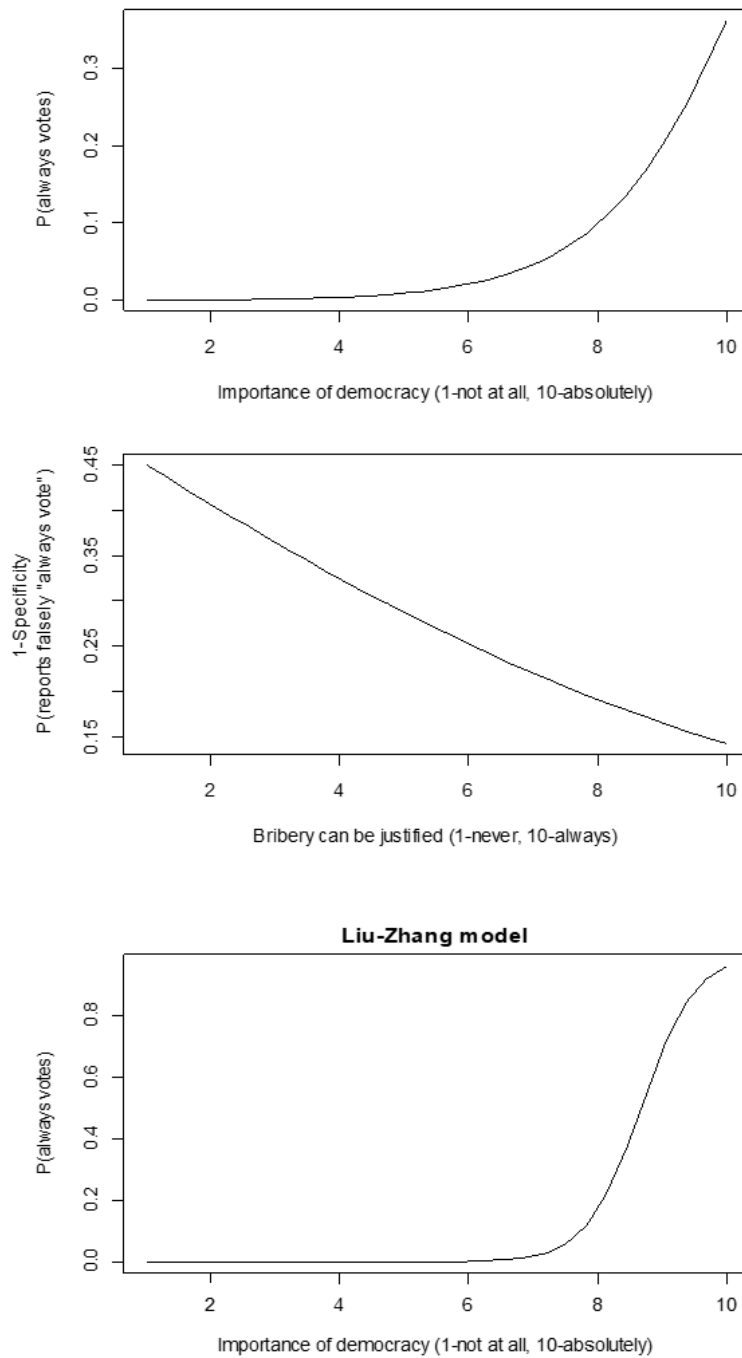


Figure 12. - Results of our model and those of the Liu-Zhang model for the Czech Republic.

5.3.3 Discussion

We extended the logistic model allowing misclassification of the outcome so that either Se or Sp may depend on some covariates while the other one is assumed to be constant. The model is a generalization of the Liu-Zhang model, a.k.a. the four-parameter logistic model.

Simulation results showed that the proposed model meets our expectations. The sample size required to reach sufficient statistical power depends on the true parameters as well as the distribution of the predictors X and Z (and hence the probability range of Y and Se).

Regarding the dependence of Y on X , the power is highest when the sensitivity varies in higher ranges, while the worst case is the opposite. It is clear why this happens so: if the Se is small, the uncertainty in the observed Y is large. If the specificity is also small, the logistic curve is very poorly estimated even for a sample of 5000 or 10000.

Power of detecting the dependence of Se on Z depends most strongly on the probability range of the outcome (Y): higher probability range results in higher power. The explanation is that sensitivity is related to the upper asymptote of the logistic curve, and this can be estimated better from data covering this region.

Interestingly, the fact that specificity is less than 1 is better detectable at Sp set to 90% than at lower Sp level. This may be explained by the higher uncertainty of estimates at lower Sp values. Since Sp affects the lower asymptote of the curve, a low outcome probability range is needed for its reliable estimation.

Application to real data illustrated the expected advantages of the model. The two examples illustrate that the model can be used for the analysis of sensitive survey questions when respondents are reluctant to answer honestly (resulting in misclassification of the outcome), and the degree of honesty is supposed to depend on certain covariates. In these situations, our model can estimate the true proportion conditional on the covariates and the degree of response bias conditional on its covariates.

However, some limitations must be kept in mind. Although our model is proven to be quite flexible it still relies heavily on the logistic relationship. The misspecification of the model may lead to strongly biased estimates of Se and Sp . Another difficulty may be a strong correlation between the linear predictors X and Z because the model might become unstable in such scenarios. A potential generalization might be that both Se and Sp depend on some covariates but we did not consider this in the present project. We feel that identifiability of such a model would require, as suggested in Section 5.2.4, a much more difficult discussion.

As a conclusion, simulation as well as applications to real data confirmed that the new model has some advantages over existing models and could become an important tool for data analysis.

6 Conclusion

This dissertation explored innovative methods and extensions for handling misclassification in statistical analyses. The research addressed critical gaps in the literature by developing both theoretical advancements and practical tools to improve the accuracy and reliability of statistical results in scenarios where misclassification may be present. Below is a summary of the key findings from each major chapter of the dissertation:

A novel profile likelihood confidence interval (PLCI) was introduced in Chapter 1 to estimate the true prevalence of a disease when sensitivity and specificity are estimated from independent validation samples. Comparative analyses showed that the PLCI demonstrated improved performance over existing methods, particularly in terms of coverage probability and interval length. By accounting for uncertainty in sensitivity and specificity, the PLCI method proved to be more robust than traditional approaches. The proposed heuristic adjustment further enhanced its accuracy in extreme cases, such as when observed prevalence or diagnostic test parameters approached their boundaries.

In Chapter 2 a computationally feasible method for constructing confidence intervals for risk differences (RD) and risk ratios (RR) adjusted for estimated sensitivity and specificity was introduced. The method successfully combined the techniques of Lang and Reiczigel for single proportion estimation with Zou and Donner's method for independent parameter differences. Simulations demonstrated that the proposed method provided reliable coverage probabilities even when sample sizes for the diagnostic parameter estimation were small. Application examples highlighted its potential for use in real-world epidemiological studies, where diagnostic tests are often imperfect.

A comprehensive investigation into the impact of misclassification on sample size requirements for one- and two-sample tests was included in Chapter 3. The research demonstrated that ignoring potential misclassification during study design or analysis can lead to biased results and might also lead to underpowered studies. To address this issue, novel sample size calculation methods were developed for both fixed and estimated sensitivity and specificity scenarios. These methods proved effective in maintaining statistical power while accounting for the variability introduced by diagnostic test imperfections.

In Chapter 4 we extended the so called promising zone method for adaptive clinical trials to scenarios involving misclassification. By incorporating the uncertainty associated with diagnostic test outcomes, the proposed approach allowed for more accurate and flexible

sample size re-estimation. Simulations confirmed that this method maintained desired power levels while optimizing resource allocation in clinical trials. The approach proved particularly useful in trials with binary endpoints based on diagnostic tests.

In Chapter 5, a new logistic regression model was introduced to handle covariate-dependent probabilities of misclassification. The model addressed challenges related to parameter identifiability and provided accurate estimates of regression coefficients. Empirical studies demonstrated that the proposed method outperformed traditional logistic regression models that did not account for misclassification. This advancement has significant implications for research in fields where covariate-dependent misclassification is common.

This dissertation tries to make important contributions to the field of statistical analysis and epidemiology. One of the key contributions is the development of new methods for handling misclassification, which represents a significant theoretical advancement and addresses critical gaps in existing literature. By proposing these innovative approaches, the research enhances the capacity of statistical models to account for diagnostic inaccuracies and other misclassification challenges. Additionally, the practical implementation of these methods as R functions provides accessible and computationally efficient tools for researchers, making it easier to adopt these techniques in practical applications. The evaluation of the proposed methods through extensive simulations and real-world application examples further validated their effectiveness and robustness. This comprehensive evaluation underscores the reliability of the methods and their potential to produce more accurate statistical results. Beyond epidemiology, the findings of this research have broader applications, extending to clinical trials, behavioral sciences, and other fields where categorical data play a central role.

The findings of this dissertation also have several practical implications. By accounting for misclassification during the study design phase, researchers can ensure adequate statistical power and reduce the risk of inconclusive results. The proposed methods provide more accurate and reliable estimates, thereby improving the validity of research conclusions. Furthermore, the extension of the promising zone method offers a flexible and resource-efficient approach to sample size re-estimation in clinical trials, optimizing trial outcomes and resource allocation.

Despite its contributions, this dissertation has certain limitations. Some methods rely on specific assumptions about the distribution of diagnostic test errors, which may not hold in all scenarios. This reliance on assumptions can limit the applicability of the methods in some cases. Additionally, while the proposed methods were implemented in R, their application may still require statistical expertise, potentially limiting their use by practitioners without a strong

statistical background. Furthermore, although the methods were validated through simulations and selected examples, further testing in diverse real-world settings is necessary to fully assess their robustness and adaptability.

Future research can build on this work in several ways. One promising avenue is the extension of the methods to handle multiclass problems, where misclassification occurs in scenarios with more than two categories. Additionally, exploring Bayesian approaches could provide more flexible modeling of uncertainty in sensitivity and specificity, offering a probabilistic framework for handling diagnostic inaccuracies. Integrating the proposed methods with machine learning models could further enhance classification performance and enable more sophisticated data analysis in complex research settings. Finally, extending the methods to handle misclassification in longitudinal studies would be a valuable advancement, enabling researchers to account for diagnostic errors over time and improve the reliability of long-term research findings.

References

- Ada, E. (2020). *EVS 2017-European Values Study 2017: Integrated Dataset*.
- Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons.
- Agresti, A., & Coull, B. A. (1998). Approximate is Better than “Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, *52*(2), 119–126.
<https://doi.org/10.1080/00031305.1998.10480550>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.
- Anderson, N. L., Grahn, R. A., Van Hoosear, K., & BonDurant, R. H. (2009). Studies of trichomonad protozoa in free ranging songbirds: Prevalence of *Trichomonas gallinae* in house finches (*Carpodacus mexicanus*) and corvids and a novel trichomonad in mockingbirds (*Mimus polyglottos*). *Veterinary Parasitology*, *161*(3–4), 178–186.
- Bailey, L., Vardulaki, K., Langham, J., & Chandramohan, D. (2005). *Introduction to epidemiology* (Vol. 237). Open University Press London.
- Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), i–8.
- Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine*, *35*(3), 325–347.
- Bernard, F., Vincent, C., Matthieu, L., David, R., & James, D. (2005). Tuberculosis and brucellosis prevalence survey on dairy cattle in Mbarara milk basin (Uganda). *Preventive Veterinary Medicine*, *67*(4), 267–281.
- Bland, M. (2015). *An introduction to medical statistics*. Oxford university press.
- Boelaert, F., Biront, P., Soumare, B., Dispas, M., Vanopdenbosch, E., Vermeersch, J., Raskin, A., Dufey, J., Berkvens, D., & Kerkhofs, P. (2000). Prevalence of bovine herpesvirus-1 in the Belgian cattle population. *Preventive Veterinary Medicine*, *45*(3–4), 285–295.
- Bollinger, C. R., & David, M. H. (1997). Modeling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association*, *92*(439), 827–835.
- Braden, B., & Caspary, W. F. (2001). Detection of *Helicobacter pylori* infection: When to perform which test? *Annals of Medicine*, *33*(2), 91–97.
- Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement error in nonlinear models* (Vol. 105). CRC press.
- Chen, Q., Galfalvy, H., & Duan, N. (2013). Effects of disease misclassification on exposure–disease association. *American Journal of Public Health*, *103*(5), e67–e73.

- Chen, Y. J., DeMets, D. L., & Gordon Lan, K. K. (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*, *23*(7), 1023–1038.
- Cheng, K., & Hsueh, H. (1999). Correcting bias due to misclassification in the estimation of logistic regression models. *Statistics & Probability Letters*, *44*(3), 229–240.
- Coughlin, S. S., Trock, B., Criqui, M. H., Pickle, L. W., Browner, D., & Tefft, M. C. (1992). The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *Journal of Clinical Epidemiology*, *45*(1), 1–7.
- Cui, L., Hung, H. J., & Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, *55*(3), 853–857.
- Davidov, O., Faraggi, D., & Reiser, B. (2003). Misclassification in logistic regression with discrete covariates. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *45*(5), 541–553.
- De Wit, J., Hage, J., Brinkhof, J., & Westenbrink, F. (1998). A comparative study of serological tests for use in the bovine herpesvirus 1 eradication programme in The Netherlands. *Veterinary Microbiology*, *61*(3), 153–163.
- Duffy, S., Warwick, J., Williams, A., Keshavarz, H., Kaffashian, F., Rohan, T., Nili, F., & Sadeghi-Hassanabadi, A. (2004). A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology & Community Health*, *58*(8), 712–717.
- Edwards, J. M., Walters, S. J., Kunz, C., & Julious, S. A. (2020). A systematic review of the “promising zone” design. *Trials*, *21*(1), 1–10.
- Egleston, B. L., Miller, S. M., & Meropol, N. J. (2011). The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. *Statistics in Medicine*, *30*(30), 3560–3572.
- Elsäßer, A., Regnstrom, J., Vetter, T., Koenig, F., Hemmings, R. J., Greco, M., Papaluca-Amati, M., & Posch, M. (2014). Adaptive clinical trial designs for European marketing authorization: A survey of scientific advice letters from the European Medicines Agency. *Trials*, *15*(1), 1–10.
- Farnham, P. G., Sansom, S. L., & Hutchinson, A. B. (2012). How much should we pay for a new HIV diagnosis? A mathematical model of HIV screening in US clinical settings. *Medical Decision Making*, *32*(3), 459–469.
- Flor, M., Weiß, M., Selhorst, T., Müller-Graf, C., & Greiner, M. (2020). Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification. *BMC Public Health*, *20*(1), 1–10.
- Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.
- Gao, P., Ware, J. H., & Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(6), 1184–1196.

- Gonçalves, L., Oliveira, M. R., Pascoal, C., & Pires, A. (2012). Sample size for estimating a binomial proportion: Comparison of different methods. *Journal of Applied Statistics - J APPL STAT*, 39, 1–21. <https://doi.org/10.1080/02664763.2012.713919>
- Gorber, S. C., Schofield-Hurwitz, S., Hardt, J., Levasseur, G., & Tremblay, M. (2009). The accuracy of self-reported smoking: A systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine and Tobacco Research*, 11(1), 12–24.
- Gordis, L. (2013). *Epidemiology e-book*. Elsevier Health Sciences.
- Grace, Y. Y. (2016). *Statistical analysis with measurement error or misclassification*. Springer.
- Greenland, S. (2008). Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. *Journal of Statistical Planning and Inference*, 138(2), 528–538.
- Greiner, M., & Gardner, I. A. (2000). Application of diagnostic tests in veterinary epidemiologic studies. *Preventive Veterinary Medicine*, 45(1–2), 43–59.
- Guidugli, F., Castro, A. A., & Atallah, Á. N. (2000). Antibiotics for preventing leptospirosis. *Cochrane Database of Systematic Reviews*, 4.
- Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Chapman and Hall/CRC.
- Hahn, A., Loderstädt, U., Frickmann, H., & Schwarz, N. G. (2019). Sparing the control arm using well-characterized diagnostic approaches—the Gart and Buck prevalence estimator for efficacy estimation in single-arm trials. *Journal of Laboratory Medicine*, 43(5), 279–281.
- Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, 35(2–3), 303–316.
- Hársfalvi, P., & Reiczigel, J. (2023). Profile likelihood confidence interval for the prevalence assessed by an imperfect diagnostic test. *Preventive Veterinary Medicine*, 214, 105886.
- Hársfalvi, P., & Singer, J. (2023). Confidence limits for risk differences and risk ratios adjusted for estimated sensitivity and specificity. *Biostatistics & Epidemiology*, 7(1), 1–13.
- Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2), 239–269.
- He, J., McClish, D. K., & Sabo, R. T. (2021). Evaluating Misclassification Effects on Single Sequential Treatment in Sequential Multiple Assignment Randomized Trial (SMART) Designs. *Statistics in Biopharmaceutical Research*, 1–8.

- Instruction for Use for the OnSite Toxo igG/igM Rapid Test.* (n.d.).
- Klepper, S., & Leamer, E. E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica: Journal of the Econometric Society*, 163–183.
- Kramps, J., Magdalena, J., Quak, J., Weerdmeester, K., Kaashoek, M., Maris-Veldhuis, M., Rijsewijk, F., Keil, G., & Van Oirschot, J. (1994). A simple, specific, and highly sensitive blocking enzyme-linked immunosorbent assay for detection of antibodies to bovine herpesvirus 1. *Journal of Clinical Microbiology*, 32(9), 2175–2181.
- Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, 62(1), 85–96.
- Lachenbruch, P. A. (1998). Sensitivity, specificity, and vaccine efficacy. *Controlled Clinical Trials*, 19(6), 569–574.
- Lachin, J. M. (2005). A review of methods for futility stopping based on conditional power. *Statistics in Medicine*, 24(18), 2747–2764.
- Lang, Z., & Reiczigel, J. (2014). Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity. *Preventive Veterinary Medicine*, 113, 13–22. <https://doi.org/10.1016/j.prevetmed.2013.09.015>
- Liu, H., & Zhang, Z. (2017). Logistic regression with misclassification in binary outcome variables: A method and software. *Behaviormetrika*, 44(2), 447–476.
- Liu, Y., Liu, J., & Zhang, F. (2013). Bias analysis for misclassification in a multicategorical exposure in a logistic regression model. *Statistics & Probability Letters*, 83(12), 2621–2626.
- Luan, X., Pan, W., Gerberich, S. G., & Carlin, B. P. (2005). Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Statistics in Medicine*, 24(14), 2221–2234.
- Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., & Sobel, J. D. (2011). Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. *Epidemiology*, 22(4), 589–597.
- Magder, L. S., & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146(2), 195–203.
- Martinková, P., & Hladká, A. (2023). *Computational aspects of psychometric methods*. Chapman & Hall/CRC.
- Mehta, C., Bhingare, A., Liu, L., & Senchaudhuri, P. (2022). Optimal adaptive promising zone designs. *Statistics in Medicine*, 41(11), 1950–1970.
- Mehta, C. R., & Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*, 30(28), 3267–3284.

- Millar, R. B. (2011). *Maximum likelihood estimation and inference: With examples in R, SAS and ADMB*. John Wiley & Sons.
- Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4), 843–855.
- Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, 58(3), 675–683.
- Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, 17(8), 873–890.
- Newton-Sánchez, O. A., de la Cruz Ruiz, M., Torres-Rojo, Y., Ochoa-Díaz-López, H., Delgado-Enciso, I., Hernández-Suárez, C. M., & Espinoza-Gómez, F. (2020). Effect of an ecosystem-centered community participation programme on the incidence of dengue. A field randomized, controlled trial. *International Journal of Public Health*, 65(3), 249–255.
- Niloofo, R., Fernando, N., de Silva, N. L., Karunanayake, L., Wickramasinghe, H., Dikmadugoda, N., Premawansa, G., Wickramasinghe, R., de Silva, H. J., & Premawansa, S. (2015). Diagnosis of leptospirosis: Comparison between microscopic agglutination test, IgM-ELISA and IgM rapid immunochromatography test. *PloS One*, 10(6), e0129236.
- Panbio™ DENGUE DUO IgM CAPTURE AND IgG CAPTURE ELISA test specifications*. (n.d.).
- Qiu, S. B., Yan, C., Zhou, D. H., Hou, J., Wang, Q. Q., Lin, Y., Fu, H. C., Zhang, J., Weng, Y. B., & Song, H. Q. (2012). High prevalence of *Trichomonas gallinae* in domestic pigeons (*Columba livia domestica*) in subtropical southern China. *African Journal of Microbiology Research*, 6(13), 3261–3264.
- Qiu, S.-F., Lian, H., Zou, G. Y., & Zeng, X.-S. (2018). Interval estimation for a proportion using a double-sampling scheme with two fallible classifiers. *Statistical Methods in Medical Research*, 27(8), 2478–2503.
- Qiu, S.-F., Poon, W.-Y., & Tang, M.-L. (2016). Confidence intervals for proportion difference from two independent partially validated series. *Statistical Methods in Medical Research*, 25(5), 2250–2273.
- Quirin, R., Rasolofo, V., Andriambololona, R., Ramboasolo, A., Rasolonavalona, T., Raharisolo, C., Rakotoaritahina, H., Chanteau, S., & Boisier, P. (2001). *Validity of intradermal tuberculin testing for the screening of bovine tuberculosis in Madagascar*.
- Reiczigel, J., Földi, J., & Ózsvári, L. (2010). Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. *Epidemiology and Infection*, 138, 1674–1678. <https://doi.org/10.1017/S0950268810000385>

- Reiczigel, J., Singer, J., & Lang, ZS. (2017). Exact inference for the risk ratio with an imperfect diagnostic test. *Epidemiology and Infection*, *145*(1), 187–193. <https://doi.org/10.1017/S0950268816002028>
- Rogan, W., & Gladen, B. (1978). Estimating Prevalence From Results of A Screening-test. *American Journal of Epidemiology*, *107*, 71–76. <https://doi.org/10.1093/oxfordjournals.aje.a112510>
- Sackett, D. L. (1979). Bias in analytic research. In *The case-control study consensus and controversy* (pp. 51–63). Elsevier.
- Spiegelman, D., Rosner, B., & Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, *95*(449), 51–61.
- Spiegelman, D., Schneeweiss, S., & McDermott, A. (1997). Measurement error correction for logistic regression models with an “alloyed gold standard.” *American Journal of Epidemiology*, *145*(2), 184–196.
- Subasinghe, S., Karunaweera, N. D., Kaluarachchi, A., Abayaweera, C. A., Gunatilake, M. H., Ranawaka, J., Jayasundara, D., & Gunawardena, G. S. A. (2011). Toxoplasma gondii seroprevalence among two selected groups of women. *Sri Lankan Journal of Infectious Diseases*, *1*(1).
- Toft, N., Jørgensen, E., & Højsgaard, S. (2005). Diagnosing diagnostic tests: Evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. *Preventive Veterinary Medicine*, *68*(1), 19–33.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859.
- Væth, M., & Skovlund, E. (2004). A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Statistics in Medicine*, *23*(11), 1781–1792.
- Vollset, S. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, *12* 9, 809–824. <https://doi.org/10.1002/sim.4780120902>
- von Euler-Chelpin, M., Lillholm, M., Vejborg, I., Nielsen, M., & Lynge, E. (2019). Sensitivity of screening mammography by density and texture: A cohort study from a population-based screening program in Denmark. *Breast Cancer Research*, *21*, 1–7.
- Waller, N. G., & Feuerstahler, L. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate Behavioral Research*, *52*(3), 350–370.

- Wilson, E. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22, 209–212.
<https://doi.org/10.1080/01621459.1927.10502953>
- Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports (1896-1970)*, 1432–1449.
- Zou, G. Y., & Donner, A. (2008). Construction of confidence limits about effect measures: A general approach. *Statistics in Medicine*, 27(10), 1693–1702.