7th International Symposium on Imprecise Probability: Theories and Applications, Innsbruck, Austria, 2011

# Partially identified prevalence estimation under misclassification using the Kappa coefficient

**Helmut Küchenhoff**
Department of Statistics
Ludwig-Maximilians-Universität
(LMU), Munich, Germany

**Thomas Augustin**
Department of Statistics
Ludwig-Maximilians-Universität
(LMU), Munich, Germany

**Anne Kunz**
Metronomia Clinical Research GmbH
Munich, Germany

## Abstract

We discuss prevalence estimation under misclassification. That is we are concerned with the estimation of a proportion of units having a certain property (being diseased, showing deviant behavior, etc.) from a random sample when the true variable of interest cannot be observed, but a related proxy variable (e.g. the outcome of a diagnostic test) is available. If the misclassification probabilities were known then unbiased prevalence estimation would be possible. We focus on the frequent case where the misclassification probabilities are unknown but two independent replicate measurements have been taken. While in the traditional precise probabilistic framework a correction from this information is not possible due to non-identifiability, the imprecise probability methodology of partial identification and systematic sensitivity analysis allows to obtain valuable insights into possible bias due to misclassification. We derive tight identification intervals and corresponding confidence regions for the true prevalence, based on the often reported kappa coefficient, which condenses the information of the replicates by measuring agreement between the two measurements. Our method is illustrated in several theoretical scenarios and in an example from oral health on prevalence of caries in children.

*Key words:* Partial Identification; Sensitivity Analysis; Prevalence Estimation; Kappa Coefficient; Misclassification; Identification Region; Ignorance Region.

## 1 Introduction

Many data in social sciences, econometrics, biometrics and epidemiology are complex in the sense that the available data at hand do not exactly convey the information one is looking for. Frequently, the variables of material interest cannot be observed directly or measured correctly, and one has to be satisfied with so-called surrogates or proxies, i.e., with somehow related, but different variables. This problem of non-ascertainability of certain ideal variables is referred to as measurement error (ME in the following) if the variables are continuous and as misclassification (MC) if they are discrete variables. If one ignores the principal difference between the ideal variables and their observable counterparts and just plugs in the surrogates instead of the ideal variables ('naive estimation'), then all the parameter estimators must be suspected to be severely biased. For the distorting effects of MC in different applications, see, e.g., [8, 23, 24, 53, 55].

In the last years there has been a considerable progress how to adjust for measurement error and misclassification in statistical models. Many correction procedures are available for consistent estimation in the presence of ME or MC, see in particular the monographs [6, 17], or, e.g., [44]. Most of those procedures are based on precise information about the process of measurement (and in complex models typically on Bayesian methods with precise priors, e.g., [43]). In the case of an additive measurement error, usually the variance of measurement error has to be known or to be estimated, e.g. by replicate measurements, to enable consistent estimation. In the presence of MC, knowledge of the conditional probabilities of correct classification, in the binary case called sensitivity and specificity, allows for general estimation procedures even in complex models; see [20] and [39] for fundamental work concerned with response misclassification and, e.g., [27, 29, 30, 59] for methods handling misclassified covariates. When no such information about ME or MC is available, identification problems arise and no consistent parameter estimation is possible. Important examples include the estimation in simple linear regression with covariate ME as well as the problem of estimating probability distributions of outcomes in the presence of MC. In this paper, we examine the latter problem in the spirit of the methodology of partial identification (e.g., [32]) and systematic sensitivity analysis (e.g., [52]).

One important example for estimating probability distributions in medical and clinical research is prevalence estimation, i.e. estimating the probability that a randomly sampled person of the population has a certain property, e.g. is diseased.[1] In the presence of MC, induced, e.g., by a medical examiner or a diagnostic tool, prevalence estimation using the relative frequency ignoring MC (naive estimation) is inconsistent. In this situation, a consistent estimator is available when the conditional probabilities of correct diagnosis (sensitivity and specificity) are known or can be estimated consistently. However, estimating sensitivity and specificity using a validation study usually relies on the availability of a correct diagnostic method (gold standard) in the validation sample. If such a gold standard method is not available, then it is usual practice to replicate measurements on the same unit to get some information on the quality of the measurement procedure. In the case of the availability of three independent measurements with identical sensitivity and specificity, it is still possible to obtain consistent estimators of prevalence; for a recent discussion, see [41]. Another scenario, where the parameters are identified, is the availability of two independent measurements with identical sensitivity and specificity in two different populations, see [46].

When only two replicate measurements in one population are available, the quality of measurement can be characterized by Cohen's kappa coefficient [9], which is based on the agreement of the replicates ("inter rater reliability"). Although there is a long discussion about the problems of using the kappa coefficient (e.g. [14, 50]), it is usually reported in those studies. However, no further correction is performed, since the resulting estimation model is not well-identified, making the derivation of a precise-valued estimator impossible. In contrast, the concept of partial identification and systematic sensitivity analysis provides valuable insights into the magnitude of the misclassification bias. We derive identification regions of the misclassification probabilities and the true prevalence, and confidence regions for the latter, additionally taking sample variation into account. In our example, we use data from a validation study, which consists of a subsample of our data to estimate kappa coefficient.

We understand our contribution as a typical example where imprecise probabilistic methodology provides powerful quantitative insights into the underlying structure, while the traditional precise approach, forced to choose between the extremes 'precise solution' or 'no solution', necessarily has to surrender.

[1] For ease of argumentation and influenced by the example from oral health discussed in Section 4, we use biometric terminology throughout the paper, without limiting the application of our results to that area.

The general methodology underlying our investigation adapts recent progress in the area of partial identification and systematic sensitivity analysis for possibly deficient data, also strongly related to the conservative handling of deficient data in imprecise probability settings (e.g., [12, 49, 58]). Up to now, such methods have been mostly applied to the case of missing or coarse data with an unknown deficiency mechanism (e.g. [33, 36], for surveys), notably with regard to missingness due to counterfactuality when analysing treatment effects (see e.g. [7, 15, 25, 35, 48]). Corresponding ideas have, for instance, been proposed in general settings in [13, 18], or more specifically to handle publication bias in meta analysis [11, 21], in the reanalysis of a public opinion survey [2] or to derive tight bounds on demand responses [4], and may provide an alternative to some neighborhood models in robust statistics ([1, Section 5]). Recently partial identification has also been applied in the context of misclassification ([19, 37]).

The paper is organized as follows. In Section 2, we deduce basic formulae for the relationship between the fundamental quantities characterizing our situation, i.e. observed prevalence, sensitivity, specificity and the kappa coefficient. From that, identification regions for the true prevalence are derived. In Section 3, sampling variability is incorporated into our estimates resulting in confidence intervals. In Section 4, we apply our findings to a data set of caries research before we conclude with a brief further discussion of our approach in Section 5.

## 2 Prevalence Estimation under Misclassification

At the beginning of this section the basic situation is described and notation and terminology are fixed (cf. also Table 1).

We address the problem of estimating the *prevalence* of a certain disease, i.e. a probability

$$p := P(Y = 1),$$

where

$$Y = \begin{cases} 1 & \text{diseased} \\ 0 & \text{not diseased} \end{cases}$$

denotes the indicator for the (true) disease status. Due to the possible presence of MC we cannot observe $Y$ directly, but instead the diagnosis of an examiner, which is denoted by

$$Y^* = \begin{cases} 1 & \text{diagnosis positive} \\ 0 & \text{diagnosis negative}\,. \end{cases}$$

The naive estimator $\frac{1}{n}\sum_{i=1}^{n} Y_i^*$ based on a simple random sample $Y_1^*, \ldots, Y_n^*$ of $Y^*$ of size $n$ is biased

and converges to $P(Y^* = 1)$. We call $p^* := P(Y^* = 1)$ the *naive prevalence* and denote the naive estimator based on the observed relative frequency by $\widehat{p^*}$.

| obs. $Y^*$ | true status $Y$ 1 | 0 | |
|---|---|---|---|
| 1 | $P(Y^* = 1\|Y = 1)$ *sens* | $P(Y^* = 1\|Y = 0)$ → false positive cases | $p^*$ |
| 0 | $P(Y^* = 0\|Y = 1)$ → false negative cases | $P(Y^* = 0\|Y = 0)$ *spec* | |
| | $p$ | | |

Table 1: Basic notions

The relationship between the true and the naive prevalence using sensitivity $sens := P(Y^* = 1|Y = 1)$ and specificity $spec := P(Y^* = 0|Y = 0)$ of the diagnosis is directly obtained from the law of total probability.

$$\begin{aligned} p^* &= P(Y^* = 1) \\ &= P(Y^* = 1|Y = 1) \cdot P(Y = 1) \\ &\quad + P(Y^* = 1|Y = 0) \cdot P(Y = 0) \\ &= p \cdot sens + (1 - p) \cdot (1 - spec) \end{aligned} \tag{1}$$

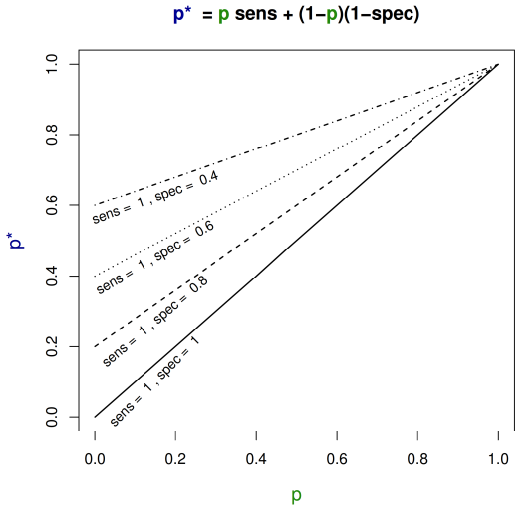


Figure 1: Illustration of misclassification bias (deviation from the angle bisector): naive (observed) prevalence $p^*$ in dependence of the true prevalence $p$ for different values of specificity and sensitivity = 1

Only for technical reasons we have to fix additionally the assumption that throughout the paper

$$sens + spec > 1\,. \tag{2}$$

This commonly used constraint is not a substantial restriction, since otherwise the diagnosis does not contain any useful information.

If sensitivity and specificity are known, equation (1) yields an unbiased estimator of $p$ by

$$\hat{p} \quad = \quad \frac{\widehat{p^*} + spec - 1}{sens + spec - 1}\,. \tag{3}$$

Moreover, Equation (1) allows to illustrate the potentially rather high distorting effects of misclassification. In Figures 1 and 2 the naive prevalence $p^*$ is plotted in dependence of the true prevalence $p$ for different misclassification probabilities. Figure 1 shows the bias in the situation of a test with optimal sensitivity, which would detect every diseased unit, but may produce a certain amount of false positive results.

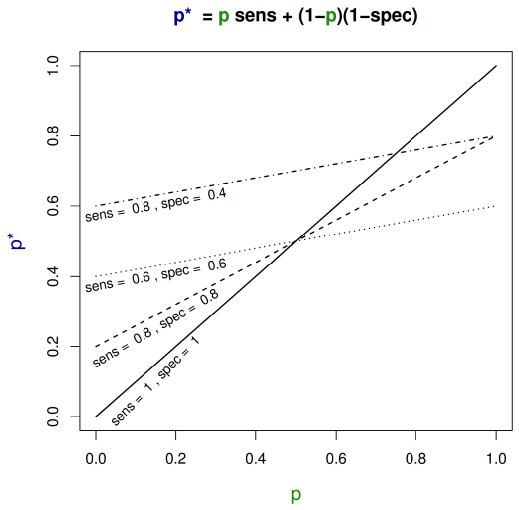


Figure 2: Illustration of misclassification bias (deviation from the angle bisector): naive (observed) prevalence $p^*$ in dependence of the true prevalence $p$ for different values of sensitivity and specificity

Figure 2 illustrates the more realistic situation of possibly false positive and false negative units. Note that in all situations the bias depends on the true, but unknown (!) value of $p$. Moreover, as in in Figure 2, the bias usually is complex in the sense that, in contrast e.g. to single-variable classical measurement error in linear regression models, even its sign can not be determined without additional knowledge. In dependence on the concrete constellation of *sense*, *spec* and $p$, over- and underestimation of the true prevalence is possible.

### 2.1 Establishing a Relationship between the Kappa Coefficient, Misclassification Probabilities and Prevalence

We now assume that we have two replicate measurements $Y_1^*, Y_2^*$ on the same units. These replicates relate to two examiners and the data can be displayed in a 2x2 table. We define the corresponding probabilities by

$$p_{jk} := P(Y_1^* = j, Y_2^* = k), \;\; i, j = 0, 1\,. \qquad (4)$$

The kappa coefficient $\kappa$ as proposed by [9], see also, e.g., [42, 45] for recent developments, assesses the chance corrected agreement among the replicate measurements (inter rater agreement). The (theoretical) kappa coefficient is defined by

$$\kappa := \frac{p_o - p_e}{1 - p_e} \qquad (5)$$

$$\begin{aligned} p_o &:= p_{00} + p_{11} \\ p_e &:= (p_{00} + p_{01}) \cdot (p_{00} + p_{10}) \\ &\quad + (p_{10} + p_{11}) \cdot (p_{01} + p_{11}) \end{aligned} \qquad (6)$$

Here, $p_o$ is the probability of the observed agreement and $p_e$ is the probability of agreement, when both ratings are unconditionally independent. The closer $\kappa$ is to 1, the better the agreement of the examiners.

**Remark 2.1** *There is an explicit relation between the kappa coefficient, the prevalence and the probabilities of misclassification, which will be useful to identify regions for the prevalence. Under the assumptions*

*(A1) Independent conditional distributions $Y_1^*|Y$ and $Y_2^*|Y$ for both replicates*

*(A2) Equal sensitivity and specificity for both replicates*

*the following equation holds ($p \in (0;1)$):*

$$\begin{aligned} \kappa &= \frac{p\,(1-p)\,(sens + spec - 1)^2}{(spec - p\,(sens + spec - 1))} \\ &\quad \cdot \frac{1}{(1 - spec + p\,(sens + spec - 1))}\,. \end{aligned} \qquad (7)$$

Equation (7) is deduced by using the assumptions (A1) and (A2) that imply

$$\begin{aligned} p_{00} &= (1-p) \cdot spec^2 + p \cdot (1 - sens)^2 \\ p_{01} &= (1-p) \cdot spec \cdot (1 - spec) \\ &\quad + p \cdot (1 - sens) \cdot sens \\ p_{10} &= p_{01} \qquad (8) \\ p_{11} &= (1-p) \cdot (1 - spec)^2 + p \cdot sens^2 \qquad (9) \end{aligned}$$

This leads, together with (5), to formula (7). Note that the kappa coefficient can be seen as a parameter of one scoring process. It is a measurement of agreement when it is independently applied on the same subject twice. It can be estimated by a validation study, where two independent scorings are available for a (sub)sample of individuals.

Note that the assumption of conditional independence and identical sensitivity and specificity may be violated, if the two replicates correspond to two different examiners, for a further discussion we refer to Section 5. The assumption of identical sensitivity and specificity can be checked using the McNemar test, which is designed for the comparison of two probabilities for dependent data. It basically checks the identity (8), see also our example in Section 4.

### 2.2 Bias Correction using the Kappa Coefficient

We want to estimate the true prevalence $p$ using the naive estimator $\widehat{p^*}$ and a given or consistently estimated kappa coefficient. The basic approach is to use equations (7) and (1) and solve them for $p$. Since there are three unknowns ($p$, *sens*, *spec*) and only two equations, there is a lack of identifiability and no direct estimator can be deduced. However, non trivial intervals $I(\vartheta \parallel p^*, \kappa)$ for the possible solutions for the three parameters $\vartheta \in \{p, sens, spec\}$ can be derived, by additionally relying on the constraint that all probabilities are in $[0;1]$. Following [32], these solutions are called identification regions. In [52] they are called ignorance regions, since they relate to ignorance in contrast to sampling error.

**Theorem 2.2** *(*Identification Regions for $p$, *sens* and *spec* using $p^*$ and $\kappa$*)*
*Let the assumptions (A1) and (A2) hold. Additionally, let $\kappa \in (0,1]$ and $sens + spec > 1$ (see (2)). Then the identification regions for the prevalence $p$, the sensitivity sens and the specificity spec based on the naive prevalence $p^* \in [0,1]$, are*

$$I(p \parallel p^*, \kappa) = \left[ \frac{p^*}{p^* + \kappa^{-1}(1-p^*)};\;\; \frac{p^*}{p^* + \kappa\,(1-p^*)} \right], \qquad (10)$$

$$I(sens \parallel p^*, \kappa) = [p^* + \kappa\,(1-p^*);\, 1] \qquad (11)$$

$$I(spec \parallel p^*, \kappa) = [1 - p^* + p^*\kappa;\, 1]\,. \qquad (12)$$

The regions in the theorem follow directly by solving equations (7) and (1), and therefore are the best that we can learn from the given values of $p^*$ and $\kappa$, without adding further assumptions. Details of the derivation are given in the web appendix ([26]).

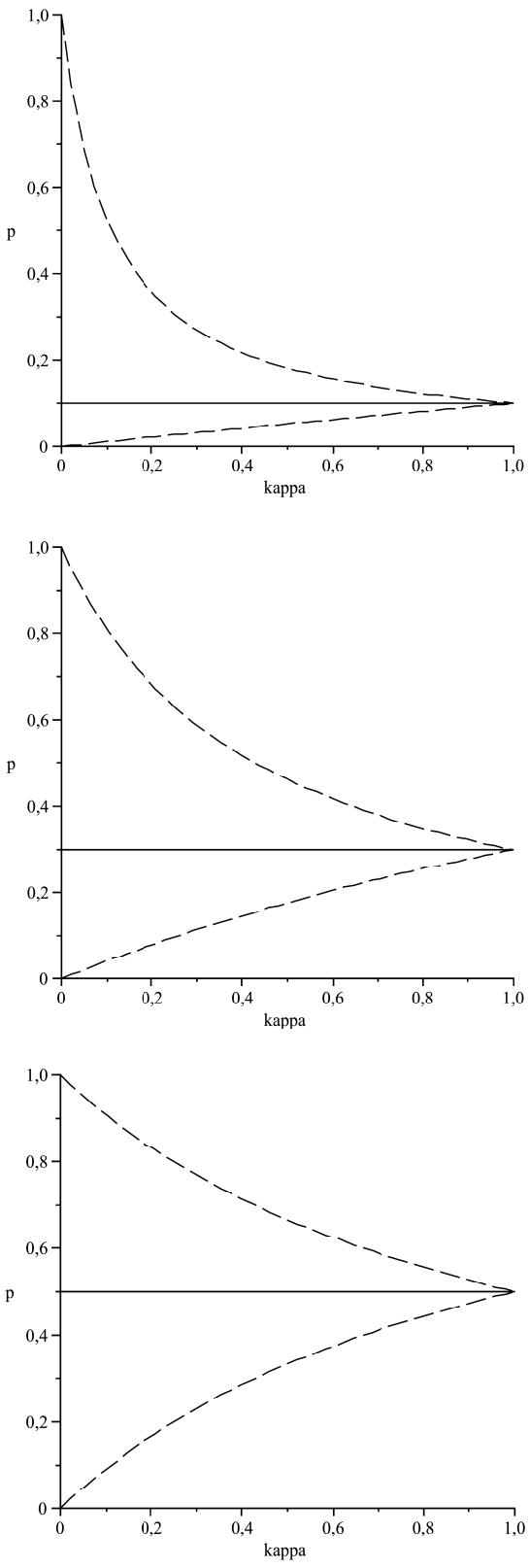


Figure 3: Identification regions (dashed lines) for the true prevalence $p$ (solid line) in dependence of $\kappa$ for different values of the naive preference $p^* \in 0.1, 0.3, 0.5$, from top to bottom.

Naturally, the width of the intervals decreases when the kappa coefficient $\kappa$ increases. Indeed, considering the extreme case where the examiners' assignments are almost random, ($\kappa \to 0$) leads to the vacuous statement $I_p = [0; 1]$. On the other hand, complete agreement, and therefore $\kappa = 1$, results in point identification, where the region for $p$ degenerates to $p^*$ and $sens = spec = 1$. In Figure 3, the identification regions are displayed as a function of the kappa coefficient for fixed values of $p^*$. For reasonable agreement of the measurements, in particular, the intervals are small enough to provide valuable insight into the true prevalence.

Note that, by construction, the method is based on the data in a conservative manner. Consequently, the identification region necessarily contains $p^*$: By $\kappa \leq 1$,

$$\begin{aligned} \frac{p^*}{p^* + \kappa^{-1}(1-p^*)} \leq \frac{p^*}{p^* + (1-p^*)} &= p^* \\ \frac{p^*}{p^* + \kappa(1-p^*)} \geq \frac{p^*}{p^* + (1-p^*)} &= p^*. \end{aligned}$$

The regions given in Theorem 2.2 are the best we can conclude from the data alone. If we interpret them as probability assignments they describe coherent interval-valued probabilities and F-probabilities in the sense of [54] and [56, 57], for details see [26]. Note that kappa coefficient and $p^*$ bear sufficient information for determining the probabilities $(p_{00}, p_{01}, p_{10}, p_{11})$, i.e. using those probabilities would not lead to an improvement of the bounds. Since the assumptions A1 and A2 imply $p_{01} = p_{10}$ and the probabilities add to 1, there are only two free parameters. An explicit formula is presented in [26].

Theorem 2.2 enables us to calculate identification regions for the prevalence, sensitivity and specificity from the naive estimator $\widehat{p^*}$ and an estimated kappa value $\hat{\kappa}$, by substituting $p^*$ and $\kappa$ with their estimators in equations (10) to (12). Note that these intervals correspond to point estimators and, in particular, are not confidence intervals. Strategies for finding confidence intervals, i.e. additionally taking the sampling variation into account, are given in the following section.

## 3 Taking Additionally Sampling Variation into Account: Confidence Intervals

We follow here the strategy from [52] and define a parameter $\gamma$, which is not identified by our data, but the other parameters of our models are identified conditional on this parameter. As a suitable choice for

this identifying parameter we propose in our context $\gamma := \frac{sens}{spec}$, which indeed would result in a point identified estimator, see (16) below. The parameter $\gamma$ has an obvious interpretation relating the probabilities of the two types of misclassification. In the framework of [52] it is called a *sensitivity* parameter. We do not use this technical term here to avoid confusion with the sensitivity of the diagnosis *sens*. The parameter $\gamma$ is restricted by (11) and (12). Therefore, the range of $\gamma$ is given by

$$[\gamma_{min},\, \gamma_{max}] = \left[p^* + \kappa\,(1-p^*), \quad \frac{1}{1-p^*+p^*\kappa}\right]. \tag{13}$$

We now assume that a consistent estimator $(\widehat{p^*}, \hat{\kappa})$ with asymptotic covariance matrix $\Sigma$ is available. If the estimator of $\kappa$ is estimated by an independent validation study, $\Sigma$ is diagonal. If we assume that $\kappa$ is known, then the corresponding entries in $\Sigma$ are 0.

To construct a confidence interval $\left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right]$ for the parameter $p$ we have to ensure that the coverage probability exceeds the confidence level $1-\alpha$ for every $\gamma \in [\gamma_{min},\, \gamma_{max}]$, i.e.

$$\inf_{\gamma \in\, [\hat{\gamma}_{min},\, \hat{\gamma}_{max}]} P_\gamma(p \in \left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right]) \geq 1-\alpha\,. \tag{14}$$

This can be achieved by defining the confidence interval as the union of confidence intervals over the identification parameter $\gamma$

$$\begin{aligned} &\left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right] := \\ &\bigcup_{\gamma \in\, [\hat{\gamma}_{min},\, \hat{\gamma}_{max}]} \left[L(\widehat{p^*}, \hat{\kappa}, \gamma); U(\widehat{p^*}, \hat{\kappa}, \gamma)\right] \end{aligned} \tag{15}$$

with $\left[L(\widehat{p^*}, \hat{\kappa}, \gamma); U(\widehat{p^*}, \hat{\kappa}, \gamma)\right]$ as suitable confidence intervals for fixed parameter $\gamma$. To calculate the latter, we apply the delta method (e.g., [3]) and use for fixed $\gamma$ the point estimator for $p$ given by

$$\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma) = \frac{(1-\widehat{p^*})\cdot\gamma - \widehat{p^*} - \sqrt{w}}{(\widehat{p^*}-1)\cdot\gamma^2 + (1-\sqrt{w})\cdot\gamma - \widehat{p^*} - \sqrt{w}} \tag{16}$$

with

$$\begin{aligned} w \;\; = \;\; & (\widehat{p^*}-1)^2\cdot\gamma^2 - 2\cdot\widehat{p^*}\cdot(\widehat{p^*}-1) \\ & \cdot(2\cdot\hat{\kappa}-1)\cdot\gamma + (\widehat{p^*})^2 \end{aligned}$$

derived from (7) and (1), see [26]. The asymptotic variance is given by the delta method

$$Var(\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma)) = D_p^T\,\Sigma\,D_p\,. \tag{17}$$

Here, $D_p$ is the vector of derivatives of $\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma)$ with respect to $\widehat{p^*}$ and $\hat{\kappa}$, and $\Sigma$ is the corresponding covariance matrix. Details are again given in [26]. Since the relationship (16) between $\gamma$ and $p$ is monotone, the choice of the confidence intervals in (15) can be optimized, see [52] or [22, 47]. If the local confidence intervals are small compared to the identification region, then it is actually justified to rely on the $(1-\alpha)\cdot 100\%$-quantile, instead of the $(1-\alpha/2)\cdot 100\%$-quantile. Thus the confidence interval is given by

$$\begin{aligned} &\left[\,L(\hat{p^*}; \hat{\kappa})\;, U(\hat{p}, \hat{\kappa})\right] = \\ &\Bigg[\hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{max}) - z_{1-\alpha}\cdot\sqrt{\widehat{Var}(\hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{max}))}; \\ &\hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{min}) + z_{1-\alpha}\cdot\sqrt{\widehat{Var}(\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{min}))}\Bigg]\,. \end{aligned} \tag{18}$$

The range for $\gamma$ is estimated using (13). Since the estimator of $(\widehat{p^*}, \hat{\kappa})$ is consistent, the probability that the interval $[\hat{\gamma}_{min},\, \hat{\gamma}_{max}]$ covers the true parameter $\gamma$ tends to 1 as sample size $n$ goes to infinity. Therefore, (15) is an asymptotic confidence interval. Note that we define our confidence intervals for the parameter and not for the entire identified set, see, in particular, [22] for a discussion of that distinction.

# 4 Example

## 4.1 The Signal-Tandmobiel® Study

| *year* | $n$ | $\widehat{p^*}$ | $se(\widehat{p^*})$ |
|---|---|---|---|
| 1996 (age 6) | 3378 | 0.118 | 0.006 |
| 1998 (age 8) | 3657 | 0.280 | 0.007 |
| 2000 (age 10) | 3415 | 0.380 | 0.008 |

Table 2: Signal-Tandmobiel® study: Estimation of $\widehat{p^*}$ per year

The Signal-Tandmobiel® study is a 6-year longitudinal oral health study, conducted in Flanders (Belgium) involving 4468 children. Data were collected on oral hygiene, gingival condition, dental trauma, prevalence and extent of enamel developmental defects, fluorosis, tooth decay, presence of restoration, missing teeth, stage of tooth eruption and orthodontic treatment need, all by using established criteria, see [51]. The children were examined annually during 1996 to 2001. Measurement of interest is the *dmft* index, which is the sum of the number of decayed, missing due to caries or filled teeth.

We use the *dmft* index as an indicator for the presence or absence of caries for each child to examine the prevalence of caries. The observed disease status $Y_i^*$

for child $i$ is

$$Y_i^* = \begin{cases} 1 & \text{caries observed} \quad (dmft > 0) \\ 0 & \text{no caries observed} \quad (dmft = 0) \end{cases}.$$

For illustration of our methods, we estimate the naive prevalence and its variance for the years 1996 (age 6), 1998 (age 8) and 2000 (age 10), see Table 2. These are the years in which a calibration study was conducted. The longitudinal structure is ignored and the naive prevalence naturally increases over the years, i.e. with the age of the children, and its standard error is very low due to the high sample size $n$.

**1996**

| | Rater 1 | | |
|---|---|---|---|
| Rater 2 | 78 | 7 | 85 |
| | 13 | 22 | 35 |
| | 91 | 29 | 120 |

$p-$value $= 0.1797$ (McNemar)
$\kappa = 0.5752(0.084)$

**1998**

| | Rater 1 | | |
|---|---|---|---|
| Rater 2 | 85 | 13 | 98 |
| | 16 | 43 | 59 |
| | 101 | 56 | 157 |

$p-$value $= 0.5775$ (McNemar)
$\kappa = 0.6023(0.066)$

**2000**

| | Rater 1 | | |
|---|---|---|---|
| Rater 2 | 89 | 14 | 103 |
| | 3 | 42 | 45 |
| | 92 | 56 | 148 |

$p-$value $= 0.0076$ (McNemar)
$\kappa = 0.7461(0.057)$

Table 3: Results of the validation study with two raters. Kappa indicates the kappa statistics with standard error in brackets.

In the calibration study in [38], the observations of the 16 regular examiners were compared to a gold standard examiner resulting in estimation of sensitivity and specificity. However, letting one single person be the gold standard examiner can still not guarantee correctness. For illustration of our methods and to incorporate this possibility of an error, the gold standard examiner is now considered a 'common' examiner. In the validation study we now have two observations per child. The results are presented in Table 3. However, since assumption A2 is questionable in our setting, we performed a McNemar test, which is based on the difference of the off diagonal cells of the two by two table. In case of the two by two table for 2000, the test indicates a significant deviation from the assumption. Therefore, we present results of our method only for the years 1996 and 1998. The estimated standard errors of the kappa coefficient are rather high due to the small sample size.

### 4.2 Correction for Misclassification

We use the methods shown in this paper to correct the estimated prevalence for misclassification. In Table 4, the corresponding identification regions based on the point estimation of $p^*$ and $\kappa$ using Theorem 2.2 are presented. The regions for the prevalence are wide. This is a consequence of the low kappa coefficient, reflecting the low agreement among the examiners. As discussed, the estimated regions include the naive estimator, but it can be seen that the naive estimator could be seriously biased. Moreover, the regions for specificity, and especially for sensitivity are wide, too.

If the kappa coefficient was considered known, the confidence intervals are only slightly smaller, indicating that the main problem is in the partial identification of our setting.

| *year* | $\widehat{p^*}$ | $\hat{\kappa}$ | $I(p \parallel \widehat{p^*}, \hat{\kappa})$ |
|---|---|---|---|
| 1996 | 0.118 | 0.577 | [0.072; 0.188] |
| 1998 | 0.280 | 0.602 | [0.190; 0.393] |
| *year* | $I(sens \parallel \widehat{p^*}, \hat{\kappa})$ | | $I(spec \parallel \widehat{p^*}, \hat{\kappa})$ |
| 1996 | [0.627; 1.000] | | [0.950; 1.000] |
| 1998 | [0.714; 1.000] | | [0.889; 1.000] |

Table 4: Signal-Tandmobiel® study: Estimated identification regions for $p$, *sens* and *spec*

In a second step, the confidence intervals for the prevalence following the strategy from Section 3 are presented in Table 5, once while incorporating the sample variability of the estimators $\widehat{p^*}$ and $\hat{\kappa}$ and, for illustration, assuming $\kappa$ to be known at its estimated value. The asymptotic confidence intervals for the naive prevalence are pretty small compared to the identification regions and to the corresponding confidence intervals, which are both based on the additional information from the kappa coefficient. Consequently, the confidence regions based on naive preva-

lence estimation still suffer from a severe overprecision. Although being somewhat large, the identification region and the corresponding confidence regions still provide valuable insight into the prevalence. For example, the hypothesis $H_0 : p \geq 0.25$ could be rejected at the 5 percent-level for the 6 year old children.

| *year* | with sampling variation of $\kappa$ | fixed $\kappa$ |
|---|---|---|
| 1996 | [0.057; 0.219] | [0.065; 0.205] |
| 1998 | [0.170; 0.416] | [0.179; 0.409] |

Table 5: Signal-Tandmobiel® study: Confidence intervals for the prevalence with and without taking the variability of $\kappa$ into account

If further nontrivial bounds on sensitivity and specificity are available by some external information, then this can be incorporated in an analogous way resulting in smaller identification regions and smaller confidence intervals based on them.

## 5 Discussion

The concept of using identification regions or intervals of ignorance in the case of misclassification with partial information on sensitivity and specificity provided by the kappa coefficient has been shown as a powerful tool for data analysis. It avoids the potentially substantial bias arising from simply ignoring misclassification if no direct correction method is available. The resulting identification regions are tight in the sense that they can not be improved without adding further assumptions. Thus they are the best that we can conclude from the data alone in this context. Our example shows that the possible effect of misclassification is rather high, even when the inter rater reliability is 'substantial´ in terms of [28]'s classification. Furthermore, the strategy of distinguishing between sampling error and ignorance due to non-identifiability is useful, since it highlights possible shortcomings in the sampling of the data structure, which cannot be compensated by a large sample size.

Since we use the value of the kappa coefficient from validation data or from other sources of information, one crucial assumption for our analysis is that this value is also correct for the main data set. This will be the case if our replication data are a random sample from our main study (internal validation). Otherwise this assumption could be disputable. It is well-known that the kappa coefficient depends on the prevalence when sensitivity and specificity are fixed [10]. So our procedure cannot be used when the prevalence in the validation data differs from the prevalence in the main study, even if we assume that the scoring procedure has fixed sensitivity and specificity. However, the latter assumption could also be problem, see the discussion in [50]. In our example, the validation study was part of a training program for the examiners. On the one hand the prevalence was higher for the validation but on the other hand there were possibly more children in that sample that were difficult to score. This could lead to values of sensitivity and specificity which are different in the main study. Nevertheless, the kappa coefficient could be nearly identical in both parts of the study. [50] performs some calculations and presents plausible scenarios for this assumption. Thus, our procedure can also be applied to studies where the value of the kappa coefficient can be transferred from the validation data to the main study even this is not true for sensitivity and specificity. Obviously, this issue has to be treated with great care.

Our results are a vivid illustration of the power of imprecise probability methods in statistical analysis based on misclassified data. As a topic of further research the conditional independence assumption (A1) in Remark 2.1 should be investigated further. As mentioned above, it may be violated if the assessments of two raters are used as substitutes for replication data, because then certain characteristics of the units may impose some dependence on the raters' judgements. We are currently studying the use of Frechet bounds and related methods in this setting. If no reliable information is available about the misclassification probabilities, our approach could be adopted to the case where sensitivity and specificity vary in certain ranges, closely relating our procedure to the 'direct method' of [37]. Then our identification parameter is two dimensional, which will result in larger identification regions.

The methodology underlying our work promises, mutatis mutandis, to be also powerful for other types of error-prone data, like misclassification of more than two categories and for (additive or multiplicative) measurement error with unknown variance. In the latter case, the availability of replicates would yield identification in many instances, but often no information about the measurement error is available, and then partially identified corrected estimators are again the best option available.

## Acknowledgements

We thank two anonymous referees for very stimulating comments, which substantially improved our paper. We thank G. Walter, M. Cattaneo and many participants of

ISIPTA'09's poster session for helpful discussions. The data collection of Signal-Tandmobiel® study was supported by Unilever, Belgium. The Signal-Tandmobiel® project comprises following partners: D. Declerck (Dental School, Catholic Univ. Leuven), L. Martens (Dental School, Univ. Ghent), J. Vanobbergen (Working Group Oral Health Promotion and Prevention, Flemish Dental Association; Dental School, Univ. Ghent), P. Pottenberg (Dental School, Univ. Brussels), E. Lesaffre (L-Biostat, Univ. Leuven, Dep. of Biostatistics, Erasmus MC, Rotterdam), K. Hoppenbrouwers (Youth Health Dep., Catholic Univ. Leuven; Flemish Assoc. for Youth Health Care).

## References

[1] T. Augustin and R. Hable. On the impact of robust statistics on imprecise probability models: A review. *Structural Safety*, 32:358–365, 2010.

[2] C. Beunckens, C. Sotto, G. Molenberghs, and G. Verbeke. A multifaceted sensitivity analysis of the Slovenian public opinion survey data. *Journal of the Royal Statistical Society: Series C*, 58(2):171–196; Corr: 575–576, 2009.

[3] P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I.* Prentice Hall, 2001.

[4] R. Blundell, M. Browning, and I. Crawford. Best nonparametric bounds on demand responses. *Econometrica*, 76(6):1227–1262, 2008.

[5] C. T. Bruckner and P. Yoder. Interpreting kappa in observational research: Baserate matters. *American Journal on Mental Retardation*, 111(6):433–441, 2006.

[6] R. Carroll, D. Ruppert, L. Stefanski, and C. Crainiceanu. *Measurement Error in Nonlinear Models.* Chapman and Hall, New York, 2006. 2nd edition.

[7] J. Cheng and D. Small. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B*, 68(5):815–836, 2006.

[8] Cheng, S., Xi, Y., and Chen, M.-H. (2008), A new mixture model for misclassification with applications for survey data, *Sociological Methods & Research*, 37, 75–104.

[9] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[10] R. Cook. Kappa and its dependence on marginal rates. In: Armitag, P., and Colton, T. (eds.) *Encyclopedia of Biostatistics.* volume 3, pp. 2166–2168. Wiley, Chichester, UK, 1998.

[11] J. Copas and D. Jackson. A bound for publication bias based on the fraction of unpublished studies. *Biometrics*, 60(1):146–153, 2004.

[12] G. De Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1-2):75–125, 2004.

[13] E. Diday and M. Noirhomme-Fraiture. Symbolic Data Analysis and the SODAS Software. Wiley and Sons, Chicester, 2008.

[14] M. Feuerman and A. Miller. Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Prctice*, 14(5):930–933, 2008.

[15] C. Gundersen and B. Kreider. Bounding the effects of food insecurity on children's health outcomes. *Journal of Health Economics*, 28(5):971–983, 2009.

[16] A. Guolo. Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*, 17(6):555–580, 2008.

[17] P. Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* Chapman and Hall, New York, 2004.

[18] P. Gustafson. Bayesian inference for partially identified models. *International Journal of Biostatistics*, 6 (2): 17, 2010.

[19] P. Gustafson and S. Greenland. Interval estimation for messy observational data. *Statistical Science*, 3 (24):328–342, 2009.

[20] J. Hausman, J. Abrevaya, and F. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2):239–269, 1998.

[21] M. Henmi, J. Copas, and S. Eguchi. Confidence intervals and p-values for meta-analysis with publication bias. *Biometrics*, 63(2):475–482, 2007.

[22] G. Imbens and C. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

[23] M. Keane and R. Sauer. Classification error in dynamic discrete choice models: implications for female labor supply behavior. *Econometrica*, 77(3):975–991, 2009.

[24] D. Kenkel, D. Lillard, and A. Mathios. Accounting for misclassification error in retrospective smoking data. *Health Economics*, 13(10):1031–1044, 2004.

[25] B. Kreider and S. Hill. Partially identifying treatment effects with an application to covering the uninsured. *Journal of Human Resources*, 44(2):409–449, 2009.

[26] H. Küchenhoff, T. Augustin, and A. Kunz. Partially identified prevalence estimation under misclassification using the Kappa coefficient (Web Appendix) www.stablab.stat.uni-muenchen.de/kuechenhoff/isipta-app.

[27] H. Küchenhoff, S. Mwalili, and E. Lesaffre. A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics*, 62(1):85–96, 2006.

[28] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[29] A. Lewbel. Estimation of average treatment effects with misclassification. *Econometrica*, 75(2):537–551, 2007.

[30] R. Lyles, A. Allen, W. Flanders, L. Kupper, and D. Christensen. Inference for case-control studies when exposure status is both informatively missing and misclassified. *Statistics in Medicine*, 25(23):4065–4080, 2006.

[31] R. Lyles, J. Williamson, H.-M. Lin, and C. Heilig. Extending McNemar's test: Estimation and inference when paired binary outcome data are misclassified. *Biometrics*, 61(1):287–294, 2005.

[32] C. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003.

[33] C. Manski. Partial identification with missing data: concepts and findings. *International Journal of Approximate Reasoning*, 39(2-3):151–165, 2005.

[34] C. Manski. Partial identification of counterfactual choice probabilities. *International Economic Review*, 48(4):1393–1410, 2007.

[35] C. Manski and J. Pepper. More on monotone instrumental variables. *Econometrics Journal*, 12(1):200–216, 2009.

[36] G. Molenberghs. Incomplete data in clinical studies: analysis, sensitivity, and sensitivity analysis. *Drug Information Journal*, 43(4):409–429, 2009.

[37] F. Molinari. Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117, 2008.

[38] S. Mwalili, E. Lesaffre, and D. Declerck. A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):77–93, 2005.

[39] J. Neuhaus. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4):843–855, 1999.

[40] J. Neuhaus. Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, 58(3):675–683, 2002.

[41] M. Pepe and H. Janes. Insights into latent class analysis of diagnostic test performance. *Biostatistics*, 8(2):474–484, 2007.

[42] C. Roberts. Modelling patterns of agreement for nominal scales. *Statistics in Medicine*, 27(6):810–830, 2008.

[43] Rummel, D., Augustin, T., & Küchenhoff, H., Correction for covariate measurement error in nonparametric longitudinal regression. *Biometrics*, 66:1209–1219, 2010.

[44] Schneeweiß, H. & Augustin, T., Some recent advances in measurement error models and methods, *ASTA Allgemeines Statistisches Archiv*, 90: 183–197, 2006.

[45] M. Shoukri and A. Donner. Bivariate modeling of interobserver agreement coefficients. *Statistics in Medicine*, 28(3):430–440, 2009.

[46] J. Stamey, D. Boese, and D. Young. Confidence intervals for parameters of two diagnostic tests in the absence of a gold standard. *Computational Statistics and Data Analysis*, 52(3):1335–1346, 2008.

[47] J. Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009.

[48] J. Stoye. Partial identification and robust treatment choice: an application to young offenders. *Journal of Statistical Theory and Practice*, 3(1):239–254, 2009.

[49] L. Utkin and T. Augustin. Decision making under imperfect measurement using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44(3):322–338, 2007.

[50] W. Vach. The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58(7):655–661, 2005.

[51] J. Vanobbergen, L. Martens, E. Lesaffre, and D. Declerck. The Signal Tandmobiel project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, 2:87–96, 2000.

[52] S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3):953, 2006.

[53] C. Vogel, H. Brenner, A. Pfahlberg, and O. Gefeller. The effects of joint misclassification of exposure and disease on the attributable risk. *Statistics in Medicine*, 24(12):1881–1896, 2005.

[54] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[55] S. Walter, C. Hsieh, and Q. Liu. Effect of exposure misclassification on the mean squared error of population attributable risk and prevented fraction estimates. *Statistics in Medicine*, 26(26):4833–4842, 2007.

[56] Weichselberger, K. (2000) The theory of interval probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24 (2-3), 149-170.

[57] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg, 2001.

[58] M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821, 2009.

[59] D. Zucker and D. Spiegelman. Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in Medicine*, 27(11):1911–1933, 2008.