LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Anne Kunz, Thomas Augustin & Helmut Küchenhoff

Partially Identified Prevalence Estimation under Misclassification using the Kappa Coefficient

Technical Report Number 074, 2010
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Partially identified prevalence estimation under misclassification using the Kappa coefficient

Anne Kunz, Thomas Augustin and Helmut Küchenhoff

January 15, 2010

Abstract

We discuss a new strategy for prevalence estimation in the presence of misclassification. Our method is applicable when misclassification probabilities are unknown but independent replicate measurements are available. This yields the kappa coefficient, which indicates the agreement between the two measurements. From this information, a direct correction for misclassification is not feasible due to non-identifiability. However, it is possible to derive estimation intervals relying on the concept of partial identification. These intervals give interesting insights into possible bias due to misclassification. Furthermore, confidence intervals can be constructed. Our method is illustrated in several theoretical scenarios and in an example from oral health, where prevalence estimation of caries in children is the issue.

Key words: partial identification; sensitivity analysis; prevalence estimation; kappa coefficient; misclassification

1 INTRODUCTION

It is well known that ignoring measurement error (ME) and misclassification (MC) may lead to severely biased parameters estimation. For the effects of the latter see e.g. (Kenkel, Lillard & Mathios 2004), (Vogel, Brenner, Pfahlberg & Gefeller 2005), (Walter, Hsieh & Liu 2007), (Keane & Sauer 2009). Many correction procedures are available for (approximately) unbiased estimation in the presence of ME or MC, see in particular the monographs (Carroll, Ruppert, Stefanski & Crainiceanu 2006, Gustafson 2004). Most of those procedures are based on some information about the process of measurement. In the case of an additive measurement error, typically the

variance of measurement error has to be known or to be estimated, e.g. by replicate measurements to enable unbiased estimation. In the presence of MC, knowledge of the conditional probabilities of correct classification, in the binary case sensitivity and specificity, allows for general estimation procedures in complex models; see (Hausman, Abrevaya & Scott-Morton 1998) and (Neuhaus 1999) for fundamental work concerned with response misclassification and, e.g. (Küchenhoff, Mwalili & Lesaffre 2006), (Lyles, Allen, Flanders, Kupper & Christensen 2006), (Lewbel 2007), (Zucker & Spiegelman 2008) 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. Well-known examples are 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; when some information about the MC process is available that is insufficient for deriving well-identified consistent estimates, but still provides valuable insight by giving non-trivial estimation regions.

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 is diseased. 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 biased. In this situation an unbiased 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 possible to obtain consistent estimators of prevalence; for a recent discussion, see (Pepe & Janes 2007). Another scenario, where the parameters are identified, is the availability of two independent measurements with identical sensitivity and specificity in two different populations, see (Stamey, Boese & Young 2008).

When only two replicate measurements in one population are available, the quality of measurement is characterized by Cohen's kappa coefficient (Cohen 1960), 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 (Vach 2005, Feuerman & Miller 2008), it is reported in many studies but no further correction is performed. This had been supported by the fact that a correction for MC only based on the kappa coefficient had been understood as being infeasible since the resulting

estimation model is not well-identified. In this paper, we develop and explore a new method for using the information on kappa coefficient to obtain valuable insights into prevalence. In the spirit of the methodology of partial identification (e.g. (Manski 2003)) and systematic sensitivity analysis (Vansteelandt, Goetghebeur, Kenward & Molenberghs 2006), we construct tight estimation intervals for the prevalence and discuss their properties.

The paper is organized as follows. In Section 2, we deduce basic formulae for the relationship between prevalence, observed prevalence, sensitivity, specificity and the kappa coefficient. This yields estimation intervals for the prevalence. 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

We address the problem of estimating the prevalence of a certain disease, i.e. a probability $p := P(Y = 1)$, where Y 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^*. The naive estimator $\frac{1}{n}\sum_{i=1}^{n} Y_i^*$ based on a simple random sample 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 $\widehat{p^*}$. The relationship between the true and the naive prevalence using sensitivity *sens* and specificity *spec* of the diagnosis is given by the following equations:

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

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

$$\hat{p} = \frac{\widehat{p^*} + spec - 1}{sens + spec - 1}. \tag{2}$$

The denominator of the equation (2) leads to the assumption

$$sens + spec > 1, \tag{3}$$

which we will require to be satisfied in the whole analysis in this paper. This commonly used assumption is not a substantial restriction, since otherwise the diagnosis does not contain any useful information.

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

The kappa coefficient κ as proposed by (Cohen 1960), see also, e.g., (Roberts 2008) and (Shoukri & Donner 2009) for recent developments, assesses the chance corrected agreement among replicate measurements on the same units. Usually the replicates correspond to different raters or examiners. The closer κ is to 1, the better the agreement of the examiners. Considering the case of two replicates (examiners) Y_1^*, Y_2^*, the (theoretical) kappa coefficient is defined by

$$\begin{aligned} \kappa &:= \frac{p_o - p_e}{1 - p_e} \\ p_{jk} &:= P(Y_1^* = j, Y_2^* = k) \\ p_o &:= p_{00} + p_{11} \\ p_e &:= (p_{00} + p_{01}) \cdot (p_{00} + p_{10}) + (p_{10} + p_{11}) \cdot (p_{01} + p_{11}) \end{aligned} \tag{4}$$

Here, p_o is the probability of the observed agreement and p_e is the expected agreement by chance.

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 ($p \in (0;1)$).

$$\kappa = \frac{p\,(1-p)\,(sens + spec - 1)^2}{(spec - p\,(sens + spec - 1)) \cdot (1 - spec + p\,(sens + spec - 1))} \tag{5}$$

Equation (5) is deduced by using the assumptions (A1) and (A2) that imply

$$\begin{aligned} p_{00} &= (1-p) \cdot spec^2 + p \cdot (1 - sens)^2 \\ p_{01} &= (1-p) \cdot spec \cdot (1 - spec) + p \cdot (1 - sens) \cdot sens \\ p_{10} &= (1-p) \cdot (1 - spec) \cdot spec + p \cdot sens \cdot (1 - sens) \\ p_{11} &= (1-p) \cdot (1 - spec)^2 + p \cdot sens^2 \end{aligned} \tag{6}$$

This leads, together with (4), to formula (5). 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.

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 (5) 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 (Manski 2003), these solutions are called identification regions. In (Vansteelandt et al. 2006) they are called ignorance regions, since they relate to ignorance in contrast to sampling error.

Theorem 2.1 *(*Identification Regions for p, *sens* and *spec* using p^* and κ*)*
Let the assumptions (A1) and (A2) hold. Additionally, let the naive prevalence $p^* \in [0, 1]$, the kappa coefficient $\kappa \in (0, 1]$ and $sens + spec > 1$ (see (3)). Then the identification regions for the prevalence p, the sensitivity sens and the specificity spec are given by

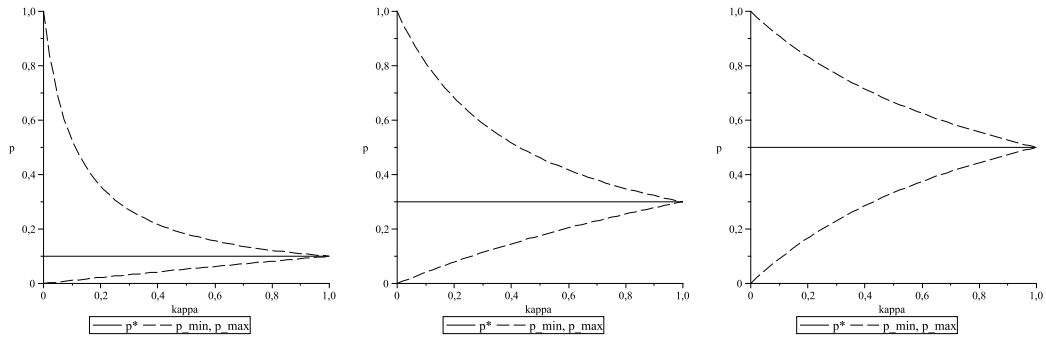
$$I(p \parallel p^*, \kappa) = \left[\frac{p^*}{p^* + \kappa^{-1}(1 - p^*)}; \ \frac{p^*}{p^* + \kappa\,(1 - p^*)} \right] \tag{7}$$

$$I(sens \parallel p^*, \kappa) = [p^* + \kappa\,(1 - p^*); \ 1] \tag{8}$$

$$I(spec \parallel p^*, \kappa) = [1 - p^* + p^*\kappa; \ 1] \tag{9}$$

The regions in the theorem follow directly by solving equations (5) and (1), and therefore are the best that we can learn from the given values of p^* and κ, without adding further assumptions. Details of the derivation are given in the appendix.

Naturally, the width of the intervals decreases when the kappa coefficient κ 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 1, 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.



(a) $p^* = 0.1$ (b) $p^* = 0.3$ (c) $p^* = 0.5$

Figure 1: Identification regions for prevalence p

Note that, by construction, the method is based on the data in a conservative manner. Consequently, the identification region necessarily contains p^*: By $\kappa \leq 1$ we conclude $\dfrac{p^*}{p^* + \kappa^{-1}(1-p^*)} \leq \dfrac{p^*}{p^* + (1-p^*)} = p^*$ and $\dfrac{p^*}{p^* + \kappa(1-p^*)} \geq \dfrac{p^*}{p^* + (1-p^*)} = p^*$.

The regions given in Theorem 2.1 are the best we can conclude from the data alone. They describe coherent interval-valued probabilities and F-probabilities in the sense of (Walley 1991) and (Weichselberger 2001), for details see Appendix C.

Theorem 2.1 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 κ with their estimators in equations (7) to (9). 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 CONFIDENCE INTERVALS

We follow here the strategy from (Vansteelandt et al. 2006) and define a parameter γ, 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 (13). The parameter γ has an obvious interpretation relating the probabilities of the two types of misclassification. In the framework of (Vansteelandt et al. 2006) 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 γ is restricted by (8) and (9). Therefore, the range of γ is given by

$$[\gamma_{min},\, \gamma_{max}] = \left[p^* + \kappa\,(1-p^*), \quad \frac{1}{1-p^*+p^*\kappa}\right]. \tag{10}$$

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

$$\inf_{\gamma \in\, [\hat{\gamma}_{min},\, \hat{\gamma}_{max}]} Prob_\gamma(p \in \left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right]) \geq 1-\alpha \tag{11}$$

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

$$\left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right] := \bigcup_{\gamma \in\, [\hat{\gamma}_{min},\, \hat{\gamma}_{max}]} \left[L(\widehat{p^*}, \hat{\kappa}, \gamma); U(\widehat{p^*}, \hat{\kappa}, \gamma)\right] \tag{12}$$

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

$$\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma) = \frac{(1-\widehat{p^*})\cdot\gamma - \widehat{p^*} - \sqrt{w}}{(\widehat{p^*}-1)\cdot\gamma^2 + (1-\sqrt{w})\cdot\gamma - \widehat{p^*} - \sqrt{w}} \tag{13}$$

$$\text{with} \qquad w = (\widehat{p^*}-1)^2\cdot\gamma^2 - 2\cdot\widehat{p^*}\cdot(\widehat{p^*}-1)\cdot(2\cdot\hat{\kappa}-1)\cdot\gamma + (\widehat{p^*})^2$$

derived from (5) and (1) according to Appendix II. The asymptotic variance is given by the delta method

$$Var(\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma)) = D_p^T\,\Sigma\,D_p \tag{14}$$

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 Σ is the corresponding covariance matrix. Details are given in Appendix II.

Since the relationship (13) between γ and p is monotone, the choice of the confidence intervals in (12) can be optimized, see (Vansteelandt et al. 2006) or (Imbens & Manski 2004), (Stoye 2009a). If the local confidence intervals are small compared to the identification region the confidence interval is given by

$$\begin{aligned}&\left[L(\hat{p^*}; \hat{\kappa})\,, U(\hat{p}, \hat{\kappa})\right] = \\ &\left[\hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{max}) - z_{1-\alpha}\cdot\sqrt{\widehat{Var}(\hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{max}))};\;\; \hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{min}) + z_{1-\alpha}\cdot\sqrt{\widehat{Var}(\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{min}))}\right]\end{aligned} \tag{15}$$

The range for γ is estimated using (10). 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 γ tends to 1 as sample size n goes to infinity. Therefore, (15) is an asymptotic confidence interval.

4 EXAMPLE

4.1 The Signal-Tandmobiel® Study

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 (Vanobbergen, Martens, Lesaffre & Declerck 2000). 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 1. 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.

Table 1: Signal-Tandmobiel® study: Estimation of $\widehat{p^*}$ per year

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

In the calibration study in (Mwalili, Lesaffre & Declerck 2005), 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. To incorporate this possibility of an error, the gold standard examiner is now considered a 'common' examiner. For illustration of our methods we assume that all examiners have the same, conditionally independent, scoring behavior (i.e. the same sensitivity and specificity), satisfying assumptions (A1) and (A2). The results for estimating the kappa coefficient are presented in Table 2. According to the classification proposed by (Landis & Koch 1977), the agreement of the observers in the years 1996 and 1998 is only moderate, but in the year 2000 the agreement is substantial. The estimated standard errors of the kappa coefficient are rather high due to the small sample size. A possible explanation for the increase of the kappa coefficient over the years is that the observers improve their examination of caries due to their experience and calibration exercises over time (Mwalili et al. 2005).

Table 2: Signal-Tandmobiel® study: Estimation of κ per year

year	n	$\hat{\kappa}$	$se(\hat{\kappa})$
1996	120	0.577	0.080
1998	157	0.602	0.066
2000	148	0.746	0.057

4.2 Correction for Misclassification

We use the methods shown in this paper to correct the estimated prevalence for misclassification. In Table 3, the corresponding identification regions based on the point estimation of p^* and κ using Theorem 2.1 are presented. The regions for the prevalence in the years 1996, 1998 and 2000 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.

Table 3: Signal-Tandmobiel® study: Estimated identification regions for p, *sens* and *spec*

year	$\widehat{p^*}$	$\hat{\kappa}$	$I(p \parallel \widehat{p^*}, \hat{\kappa})$	$I(sens \parallel \widehat{p^*}, \hat{\kappa})$	$I(spec \parallel \widehat{p^*}, \hat{\kappa})$
1996	0.118	0.577	[0.072; 0.188]	[0.627; 1.000]	[0.950; 1.000]
1998	0.280	0.602	[0.190; 0.393]	[0.714; 1.000]	[0.889; 1.000]
2000	0.380	0.746	[0.314; 0.451]	[0.843; 1.000]	[0.903; 1.000]

(Mwalili et al. 2005) have interpreted the validation study in another way. In Table 4 the results for the standard matrix method correction and the corresponding confidence intervals are given. There it can be seen that the corresponding corrected estimator for 1996 is not in the estimated identification region. This highlights the problem and the relevance of the correct interpretation of the validation study.

Table 4: Signal-Tandmobiel® study: Corrected prevalence estimation with matrix method

year	$\widehat{p^*}$	$\hat{sens}$	$\hat{spec}$	$\hat{p}$	CI
1996	0.118	0.629	0.918	0.094	[0.007; 0.182]
1998	0.280	0.729	0.867	0.247	[0.147; 0.347]
2000	0.380	0.933	0.864	0.306	[0.239; 0.373]

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 κ to be known at its estimated value.

Table 5: Signal-Tandmobiel® study: Confidence intervals for estimated regions

year	$\widehat{p^*}$	$\hat{\kappa}$	$[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})]$	$[L(\widehat{p^*}, \kappa); U(\widehat{p^*}, \kappa)]$
1996	0.118	0.577	[0.057; 0.219]	[0.065; 0.205]
1998	0.280	0.602	[0.170; 0.416]	[0.179; 0.409]
2000	0.380	0.746	[0.297; 0.468]	[0.300; 0.467]

If the kappa coefficient was considered known, the confidence intervals are only slightly smaller, although the sampling variability for $\widehat{p^*}$ is rather low due to the high sample size in the validation study.

Finally, the following figure graphically summarizes the different methods estimating the prevalence. 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 prevalence 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 $H_0 : p \leq 0.25$ could be rejected at the 5

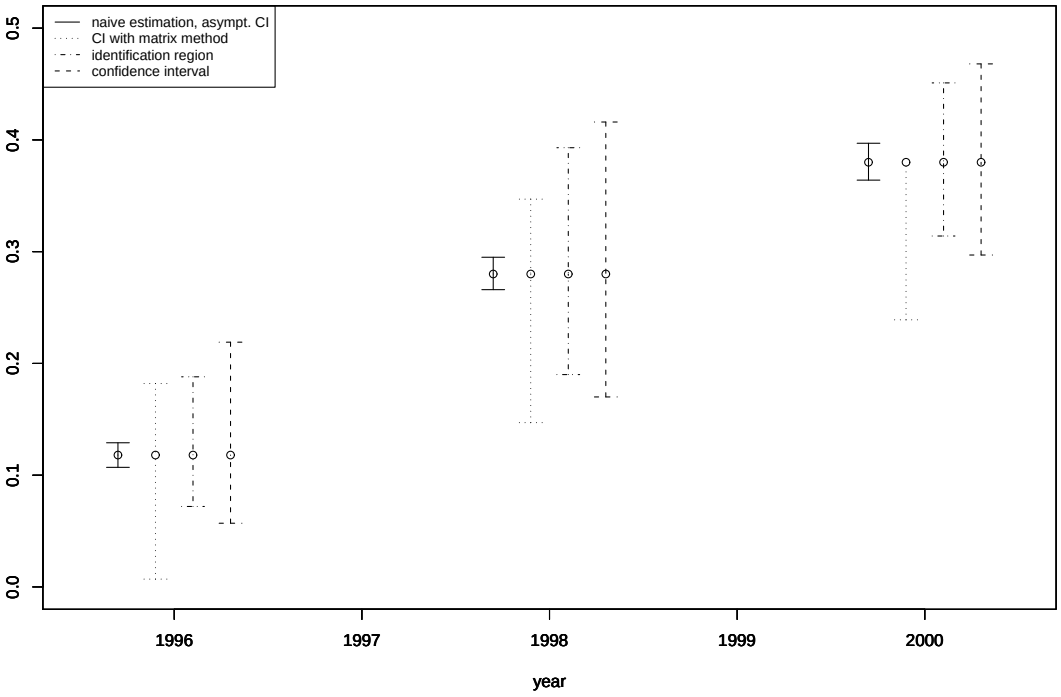


Figure 2: Signal-Tandmobiel®: Prevalence estimation: identification region and confidence limits

percent-level for the 10 year old children.

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 assumption and so 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 (Landis & Koch 1977). 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 (Cook 1998). 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 (Vach 2005). In our example, the validation study was part of a training programm 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.

However, the kappa coefficient could be nearly identical in both parts of the study. (Vach 2005) 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.

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 (Molinari 2008). Then our identification parameter is two dimensional, which will result in larger identification regions.

The general methodology underlying our investigation can also be seen as a systematic sensitivity analysis for possibly deficient data, also strongly related to the conservative handling of deficient data in imprecise probability settings (e.g. (De Cooman & Zaffalon 2004), (Utkin & Augustin 2007), (Zaffalon & Miranda 2009)). Up to now, such methods have been mostly applied to the case of missing data with an unknown missing mechanism (e.g. (Manski 2005), (Molenberghs 2009), for surveys), notably with regard to missingness due to counterfactuality when analysing treatment effects (see e.g. (Cheng & Small 2006), (Gundersen & Kreider 2009), (Kreider & Hill 2009), (Stoye 2009b), (Manski & Pepper 2009)). Related ideas have, for instance, been applied to handle publication bias in meta analysis (Copas & Jackson 2004, Henmi,

Copas & Eguchi 2007), in the reanalysis of a public opinion survey (Beunckens, Sotto, Molenberghs & Verbeke 2009) or to derive tight bounds on demand responses (Blundell, Browning & Crawford 2008).

Our approach promises, mutatis mutandis, to also be powerful for other types of error-prone data, like misclassification for 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 the best option available.

Acknowledgement

We thank Gero Walter for very 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 University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Working Group Oral Health Promotion and Prevention, Flemish Dental Association; Dental School, University Ghent), P. Pottenberg (Dental School, University Brussels), E. Lesaffre (L-Biostat, University Leuven, Department of Biostatistics, Erasmus MC, Rotterdam), K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

A Proof of Theorem 2.1

Proof of Theorem 2.1 Assume $p^* \in (0;1)$ and $\kappa \in (0;1)$ and let the assumptions (3), (A1) and (A2) hold. To proof the theorem the set of equations´ (see equations (1) and (5))

$$\begin{aligned} p^* &= p \cdot sens + (1-p) \cdot (1 - spec) \\ \kappa &= \frac{p\,(1-p)\,(sens + spec - 1)^2}{(spec - p\,(sens + spec - 1)) \cdot (1 - spec + p\,(sens + spec - 1))} \end{aligned} \tag{16}$$

is required. It consists of the relation of p, *sens* and *spec* to p^* and to κ, respectively. For the proof, the following Lemmata are useful.

1. **Lemma A.1** *The maximal values $sens_{max}$ and $spec_{max}$ for sensitivity and specificity based on the set of equations (16) are*

$$sens_{max} = 1 \qquad and \qquad spec_{max} = 1$$

This is justified by substituting $sens = 1$ in (16), which leads to the unique solution

$$p = p^* \left[p^* + \kappa^{-1} (1 - p^*) \right]^{-1} \in (0;1) \qquad and \qquad spec = 1 - p^* + p^* \kappa \in (0;1)$$

and substituting $spec = 1$, which leads to

$$p = p^* \left[p^* + \kappa (1 - p^*) \right]^{-1} \in (0;1) \qquad and \qquad sens = p^* + \kappa (1 - p^*) \in (0;1).$$

2. **Lemma A.2** *(a) Assume sens is fixed. Then p is increasing with increasing spec.*
(b) Assume spec is fixed. Then p is decreasing with increasing sens and vice versa.
(c) Assume p is fixed. Then sens is decreasing with increasing spec and vice versa.

As a first step to prove this lemma note that with assumption (3) $sens + spec > 1 \iff 1 - spec < sens$ and equation (1) we get

$$p^* < p \cdot sens + (1-p) \cdot sens = sens \tag{17}$$

and analogously

$$p^* > 1 - spec. \tag{18}$$

Then the relation of p, *sens* and *spec* based on the set of equations (16) is analyzed, where at any time one of these parameters is assumed to be fixed, which leads to the statements 2.(a)-(c). In detail the following relationships arise:

(a) Assume $sens$ is fixed. Then the relation of p and $spec$ is

$$p(spec) = \frac{(1-p^*-spec)^2}{(1-p^*-spec)^2+\kappa\, p^*\,(1-p^*)}$$

$$\frac{\partial\, p(spec)}{\partial\, spec} = \frac{2\,\kappa\, p^*\,(1-p^*)\,(p^*+spec-1)}{((1-p^*-spec)^2+\kappa\, p^*\,(1-p^*))^2} > 0 \qquad \text{since, according to (18),} \quad spec > 1-p^*$$

(b) Assume $spec$ is fixed. Then the relation of p and $sens$ is

$$p(sens) = \frac{\kappa\, p^*\,(1-p^*)}{\kappa\, p^*\,(1-p^*)+(p^*-sens)^2}$$

$$\frac{\partial\, p(sens)}{\partial\, sens} = \frac{2\,\kappa\, p^*\,(1-p^*)\,(p^*-sens)}{(\kappa\, p^*\,(1-p^*)+(p^*-sens)^2)^2} < 0 \qquad \text{since, in the light of (17),} \quad sens > p^*$$

(c) Assume p is fixed. Then the relation of $sens$ and $spec$ is

$$sens(spec) = \frac{(1-p^*-spec)\,p^*-\kappa\, p^*\,(1-p^*)}{1-p^*-spec} \tag{19}$$

$$\frac{\partial\, sens(spec)}{\partial\, spec} = \frac{\kappa\, p^*\,(p^*-1)}{(1-p^*-spec)^2} < 0 \qquad \text{since, by (18), in particular,} \quad spec \neq 1-p^*$$

Using Lemma A.1 and Lemma A.2, the limits $\underline{I}(p \parallel p^*,\kappa)$ and $\overline{I}(p \parallel p^*,\kappa)$ of $I(p \parallel p^*,\kappa)$ are solutions of the set of equations (16), calculated as follows: Inserting the maximal value $\overline{I}(sens \parallel p^*,\kappa) = 1$ and solving (16) leads to $\underline{I}(p \parallel p^*,\kappa)$ and inserting $\overline{I}(spec \parallel p^*,\kappa) = 1$ to $\overline{I}(p \parallel p^*,\kappa)$, see formula (7).

For $p^* = 0$, $p^* = 1$ and $\kappa = 1$ the identification regions are not defined, but continuously continuable.

The calculation of I_{sens} and I_{spec} is analogous.

B Details of Section 3

With $\gamma = \frac{sens}{spec} > 0$ the set of equations (16) now has the form

$$\begin{aligned} p^* &= p\cdot\gamma\cdot spec + (1-p)\cdot(1-spec) \\ \kappa &= \frac{p\,(1-p)\,(\gamma\cdot spec+spec-1)^2}{(spec-p\,(\gamma\cdot spec+spec-1))\cdot(1-spec+p\,(\gamma\cdot spec+spec-1))} \end{aligned}$$

which leads to the following formulas for p and $spec$ in dependence of γ

$$\begin{aligned} p(\gamma) &= -\frac{(p^*-1)\,\gamma+p^*+\sqrt{w}}{(p^*-1)\,\gamma^2+(1-\sqrt{w})\,\gamma-p^*-\sqrt{w}} \\ spec(\gamma) &= 0.5\,(1-p^*+\gamma^{-1}\,(p^*+\sqrt{w})) \end{aligned}$$

Using the delta method (e.g. (Bickel & Doksum 2001)), the asymptotic variance of the corresponding estimator $\widehat{p}(\gamma)$ of the prevalence p in dependence of γ is given by

$$\begin{aligned} var(\widehat{p}(\gamma)) &= \frac{16\,\gamma^4}{(\gamma - p^* - \sqrt{w} + \gamma^2\,(p^*-1) - \sqrt{w}\,\gamma)^4\,w}\cdot \\ &\Big[(1+2\,\gamma+\gamma^2)\,var(\hat{\kappa})\,p^{*6} - (2+6\,\gamma+4\,\gamma^2)\,var(\hat{\kappa})\,p^{*5} \\ &+(1+6\,\gamma+6\,\gamma^2)\,var(\hat{\kappa})\,p^{*4} - (4\,\gamma^2+2\,\gamma)\,var(\hat{\kappa})\,p^{*3} \\ &+(\kappa^2\,var(\widehat{p^*})\,\gamma^2 + \kappa^2\,var(\widehat{p^*}) - 2\,\kappa^2\,var(\widehat{p^*})\,\gamma + var(\hat{\kappa})\,\gamma^2)\,p^{*2} \\ &+\,(-2\,\kappa^2\,var(\widehat{p^*})\,\gamma^2 + 2\,\kappa^2\,var(\widehat{p^*})\,\gamma)\,p^* + \kappa^2\,var(\widehat{p^*})\,\gamma^2\Big] \end{aligned}$$

with

$$w = (p^*-1)^2\cdot\gamma^2 - 2\cdot p^*\cdot(p^*-1)\cdot(2\cdot\kappa-1)\cdot\gamma + p^{*2}$$

and the asymptotic variance of *spec* is given by

$$var(\widehat{spec(\gamma)}) = \frac{((1-p^*)\,\gamma^2 + (\sqrt{w} + 4\,\kappa\,p^* - 2\,\kappa - 2\,p^* + 1)\,\gamma - \sqrt{w} - p^*)^2\,var(\widehat{p^*})}{4\,w\,\gamma^2} + \frac{p^{*2}\,(p^*-1)^2\,var(\hat{\kappa})}{w}$$

C Recoding and identification regions

We provide a short proof that identification regions are logically consistent in the sense that recoding does not change the conclusions to be drawn. Denote after transition from Y to $\widetilde{Y} := 1-Y$ and Y^* to $\widetilde{Y}^* := 1-Y^*$ the corresponding prevalence by $\widetilde{p} := P(\widetilde{Y}=1) = 1-p$ and the corresponding naive prevalence by $\widetilde{p}^* = P(\widetilde{Y}^*=1) = 1-p^*$, respectively, and note that sensitivity and specificity exchange their role, $\widetilde{sens} = P(\widetilde{Y}^*=1 \mid \widetilde{Y}=1) = P(Y^*=0 \mid Y=0) = spec$ and $\widetilde{spec} = P(\widetilde{Y}^*=0 \mid \widetilde{Y}=0) = P(Y^*=1 \mid Y=1) = sens$, while κ is not changed by this recoding procedure. i.e. $\kappa = \tilde{\kappa}$:

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

Similar algebra shows that the corresponding identification regions are conjugated., i.e. the lower bound of $I(p \parallel p^*, \kappa)$ and the upper bound of $I(\tilde{p} \parallel \tilde{p}^*, \tilde{\kappa})$ as well as the upper bound of $I(p \parallel p^*, \kappa)$ and the lower bound of $I(\tilde{p} \parallel \tilde{p}^*, \tilde{\kappa})$ sum up to one, and thus, since Y is dichotomous, $I(p \parallel p^*, \kappa)$ and $I(\tilde{p} \parallel \tilde{p}^*, \tilde{\kappa})$ describe coherent interval-valued probabilities and F-probabilities in the sense of (Walley 1991) and (Weichselberger 2001). Moreover, the identification regions of *sens* and $\widetilde{spec}$, as well as of *spec* and $\widetilde{sens}$, coincide.

References

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

Bickel, P. & Doksum, K. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I*, Prentice Hall.

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

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

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

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

Cook, R. (1998). Kappa and its dependence on marginal rates, Vol. 3, Wiley, Chichester, UK, pp. 2166–2168. Encyclopedia of Biostatistics.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Manski, C. (2003). *Partial Identification of Probability Distributions*, Springer, New York.

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

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

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

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

Mwalili, S., Lesaffre, E. & Declerck, D. (2005). 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.

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

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

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

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

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

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

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

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

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

Vanobbergen, J., Martens, L., Lesaffre, E. & Declerck, D. (2000). 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.

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

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

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London.

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

Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*, Physika, Heidelberg.

Zaffalon, M. & Miranda, E. (2009). Conservative inference rule for uncertain reasoning under incompleteness, *Journal of Artificial Intelligence Research* **34**: 757–821.

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