Partially Identified Prevalence Estimation under Misclassification using the Kappa-Coefficient

Anne Kunz
Munich Cancer Registry
IBE, Biometry and Bioinformatics, University of Munich (LMU)
anne.kunz@gmx.net

Thomas Augustin
Department of Statistics, University of Munich (LMU)
thomas@stat.uni-muenchen.de

Helmut Küchenhoff
Department of Statistics, University of Munich (LMU)
kuechenhoff@stat.uni-muenchen.de

Anne Kunz, Thomas Augustin, Helmut Küchenhoff, LMU München
1. Misclassification

• Interest in

\[ Y_i = \begin{cases} 
1 & \text{diseased} \\
0 & \text{not diseased} 
\end{cases} \]

\[ p := \mathbb{P}(Y = 1) \text{ prevalence} \]

• Often only available

\[ Y_i^* = \begin{cases} 
1 & \text{test positive} \\
0 & \text{test negative} 
\end{cases} \]

\[ p^* := \mathbb{P}(Y^* = 1) \text{ naive prevalence} \]

• \( Y_i^* \leftarrow Y_i \) misclassification
The Signal-Tandmobiel® Study

- 6 years longitudinal oral health study (1996 - 2001)
- 4468 children in Flanders (Belgium)
- Annual examinations
- Presence/absence of caries

\[ Y_i^* = \begin{cases} 
1 & \text{caries observed} \\
0 & \text{no caries observed} 
\end{cases} \quad \leftrightarrow \quad Y_i = \begin{cases} 
1 & \text{caries} \\
0 & \text{no caries} 
\end{cases} \]
## 2. Misclassification Bias

<table>
<thead>
<tr>
<th>$Y^* = 1$</th>
<th>$Y = 1$</th>
<th>$Y = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{P}(Y^* = 1</td>
<td>Y = 1)$</td>
<td>$\mathbb{P}(Y^* = 1</td>
</tr>
<tr>
<td><strong>sensitivity</strong></td>
<td><strong>false positive</strong></td>
<td></td>
</tr>
<tr>
<td>$Y^* = 0$</td>
<td>$\mathbb{P}(Y^* = 0</td>
<td>Y = 1)$</td>
</tr>
<tr>
<td><strong>false negative</strong></td>
<td><strong>specificity</strong></td>
<td></td>
</tr>
</tbody>
</table>

Assume throughout reasonable quality of the test: $sens + spec > 1$. \(1\)

Prevalence estimation based on misclassified data may be severely biased.

Anne Kunz, Thomas Augustin, Helmut Küchenhoff, LMU München
\[ p^* = p \cdot \text{sens} + (1 - p) \cdot (1 - \text{spec}) = \]
\[ = p \cdot \text{sens} + 1 - \text{spec} - p = \]
\[ = p (\text{sens} + \text{spec} - 1) + (1 - \text{spec}) \]
Figure 1: Illustration of misclassification bias (deviation from the angle bisector).
3. Correcting for Misclassification

3.1 The Extreme Cases

- If \( sens \) and \( spec \) are precisely known, (2) yields (together with (1)) an unbiased corrected prevalence estimator

\[
\hat{p} = \frac{\hat{p}^* + spec - 1}{sens + spec - 1}.
\]  

- If there is complete ignorance on \( sens \) and \( spec \) then the corrected prevalence estimator is vacuous:

\[
\hat{p} = [0; 1].
\]
3.2 Use Additional Knowledge: kappa Coefficient

- Quite often two replicated measurements (two raters)
- In particular in medicine, common characterization of the quality of measurements by kappa, a coefficient of inter-rater agreement

<table>
<thead>
<tr>
<th>$Y^*(2)$</th>
<th>$Y^*(1)$</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_{11}$</td>
<td>$p_{10}$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>$p_{01}$</td>
<td>$p_{00}$</td>
<td></td>
</tr>
</tbody>
</table>

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

with

$$p_{jk} = P(Y^*(1) = j, Y^*(2) = k), \quad p_o = p_{00} + p_{11}$$

$$p_e = (p_{00} + p_{01}) \cdot (p_{00} + p_{10}) + (p_{10} + p_{11}) \cdot (p_{01} + p_{11})$$
Classification of $\kappa$ according to Landis & Koch

Landis and Koch (1977) proposed a widely used classification of the kappa-statistic (see Table 1) to "maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics [...]". Although these divisions are clearly arbitrary, they do provide useful 'benchmarks' for the discussion".

Table 1: Classification of $\kappa$ according to Landis & Koch (and alternative common terminology)

<table>
<thead>
<tr>
<th>kappa-statistic</th>
<th>strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 0.00$</td>
<td>poor</td>
</tr>
<tr>
<td>0.01 - 0.20</td>
<td>slight (insufficient)</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>fair (satisfactory)</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>moderate (sufficient)</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>substantial (good)</td>
</tr>
<tr>
<td>0.81 - 1.00</td>
<td>almost perfect (excellent)</td>
</tr>
</tbody>
</table>
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 can be deduced:

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

Together with (see (2) above)

$$
p^* = p \cdot sens + (1 - p) \cdot (1 - spec)
$$

this does not yield a unique solution.
3.3 Point Identification through Additional Constraints

Traditional way to proceed: add additional assumptions leading to point identification: \( sens = spec \), or more general by \( \frac{sens}{spec} =: \gamma \) known.

Then the system described by (4) and (5) indeed has a unique solution, leading to

\[
p(\gamma, \hat{p}^*, \hat{\kappa}) = \frac{(1 - p^*) \cdot \gamma - p^* - \sqrt{w}}{(p^* - 1) \cdot \gamma^2 + (1 - \sqrt{w}) \cdot \gamma - p^* - \sqrt{w}}
\]  

(6)

with \[ w = \sqrt{(\hat{p}^* - 1)^2 \cdot \gamma^2 - 2 \cdot \hat{p}^* \cdot (\hat{p}^* - 1) \cdot (2 \cdot \hat{\kappa} - 1) \cdot \gamma + (\hat{p}^*)^2} \]
3.4 Partial Identification – Basic Ideas

- Point identification only possible under non-testable assumptions

- Manski’s (2003) law of decreasing credibility: The credibility of inference decreases with the strength of the assumptions maintained.

- Look at empirical evidence alone, consider all models compatible with the data.

- Parallel development in econometrics partial identification (initiated by Manski) and in biometrics systematic sensitivity analysis (e.g. Vansteelandt et al. (2006, Stat. Sinica))
3.5 Identification Regions for Prevalence Estimations

Let $\kappa > 0$ and (A1), (A2) and (1) be satisfied. Then the identification regions $I(z||p^*, \kappa) = \{z \text{ satisfying (2) and (5) for given } p^*, \kappa\}$, for $z \in \{p, sens, spec\}$ are

$$I(p||p^*, \kappa) := \left[ \frac{p^*}{p^* + \kappa^{-1}(1 - p^*)}; \frac{p^*}{p^* + \kappa(1 - p^*)} \right]$$ \hspace{1cm} (7)

$$I(sens||p^*, \kappa) := [p^* + \kappa(1 - p^*); 1]$$ \hspace{1cm} (8)

$$I(spec||p^*, \kappa) := [1 - p^* + p^*\kappa; 1]$$ \hspace{1cm} (9)

$\kappa \rightarrow 0 : I_p(p^*, \kappa) = [0, 1]$

$\kappa \rightarrow 1 : I_p(p^*, \kappa) = p^*, \quad sens = spec = 1$
Figure 2: Identification regions for prevalence $p$

(a) $p^* = 0.1$

(b) $p^* = 0.3$

(c) $p^* = 0.5$
4. Confidence Intervals for Partially Identified Prevalence Estimation

- Identification regions reflect lack of knowledge ("ignorance"), but do not take into account sample variability ("uncertainty").


- Identification parameter $\gamma := \frac{\text{sens}}{\text{spec}}$; given $\gamma$ all parameters would be identified (cp. section 3.3) \(^1\)

\(^1\)Vansteelandt et. al. (2006) use the term sensitivity parameter, which, however, is not used here to avoid terminological conflict with the sensitivity sens.
• \( \gamma \) can be shown to vary between \( \gamma_{min} \) and \( \gamma_{max} \), where

\[
[\gamma_{min}, \gamma_{max}] = \left[ \kappa + p^* - \kappa \cdot p^*; \frac{1}{\kappa \cdot p^* - p^* + 1} \right].
\]  

(10)

• Construct confidence interval \([L(\hat{p}^*, \hat{\kappa}); U(\hat{p}^*, \hat{\kappa})]\) such that

\[
\inf_{\gamma \in [\hat{\gamma}_{min}, \hat{\gamma}_{max}]} Pr_{\gamma}(p \in [L(\hat{p}^*, \hat{\kappa}); U(\hat{p}^*, \hat{\kappa})]) \geq 1 - \alpha
\]

(11)

• Use suitable confidence intervals \([L(\hat{p}^*, \hat{\kappa}, \gamma); U(\hat{p}^*, \hat{\kappa}, \gamma)]\) given \( \gamma \) and take

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

(12)
Consider for fixed $\gamma$ the point estimator for $p$ derived from (6)

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

(13)

with 

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

The asymptotic variance is given by the delta method

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

(14)

with, $D_p$ as the vector of derivatives of $\hat{p}(\gamma, \hat{p}^*, \hat{\kappa})$ with respect to $\hat{p}^*$ and $\hat{\kappa}$, and $\Sigma$ the covariance matrix of $\hat{p}^*$ and $\hat{\kappa}$. 
Since the relationship (13) between $\gamma$ and $p$ is monotone, the choice of the confidence intervals in (12) can be improved. If all the confidence intervals given $\gamma$ are small compared to the uncertainty region, an asymptotic confidence interval is given by

$$\left[ \hat{p}(\hat{\gamma}_{\text{max}}) - z_{1-\alpha} \cdot \sqrt{\hat{\text{Var}}(\hat{p}(\hat{\gamma}_{\text{max}}))}; \hat{p}(\hat{\gamma}_{\text{min}}) + z_{1-\alpha} \cdot \sqrt{\hat{\text{Var}}(\hat{p}(\hat{\gamma}_{\text{min}}))} \right]$$

(15)

$sens + spec > 1$

Since $\hat{\gamma}_{\text{min}}, \hat{\gamma}_{\text{max}}$ estimate $\gamma_{\text{min}}$ and $\gamma_{\text{max}}$ consistently, this is an asymptotic level $(1 - \alpha)$--confidence interval.
5. Results from the Signal-Tandmobiel® Study

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

<table>
<thead>
<tr>
<th>year</th>
<th>$n$</th>
<th>$\hat{p}^*$</th>
<th>$se(\hat{p}^*)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996 (age 6)</td>
<td>3378</td>
<td>0.118</td>
<td>0.006</td>
</tr>
<tr>
<td>1998 (age 8)</td>
<td>3657</td>
<td>0.280</td>
<td>0.007</td>
</tr>
<tr>
<td>2000 (age 10)</td>
<td>3415</td>
<td>0.380</td>
<td>0.008</td>
</tr>
</tbody>
</table>

- For illustration interpret validation measurement as a replicate,
- additionally satisfying (A1) and (A2)
- Further assumption: treat validation sample as a random sample
### Table 3: Signal-Tandmobiel® study: Estimation of $\kappa$ per year

<table>
<thead>
<tr>
<th>year</th>
<th>$n$</th>
<th>$\hat{\kappa}$</th>
<th>se($\hat{\kappa}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>120</td>
<td>0.575</td>
<td>0.084</td>
</tr>
<tr>
<td>1998</td>
<td>157</td>
<td>0.602</td>
<td>0.066</td>
</tr>
<tr>
<td>2000</td>
<td>148</td>
<td>0.746</td>
<td>0.057</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>year</th>
<th>$\hat{p}^*$</th>
<th>$\hat{\kappa}$</th>
<th>$\hat{p}$ min max</th>
<th>sens min max</th>
<th>spec min max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>0.118</td>
<td>0.575</td>
<td>0.072 0.189</td>
<td>0.625 1.000</td>
<td>0.950 1.000</td>
</tr>
<tr>
<td>1998</td>
<td>0.280</td>
<td>0.602</td>
<td>0.190 0.393</td>
<td>0.714 1.000</td>
<td>0.889 1.000</td>
</tr>
<tr>
<td>2000</td>
<td>0.380</td>
<td>0.746</td>
<td>0.314 0.451</td>
<td>0.843 1.000</td>
<td>0.903 1.000</td>
</tr>
</tbody>
</table>
Figure 3: Signal-Tandmobiel®: confidence limits for $\hat{p}$

Anne Kunz, Thomas Augustin, Helmut Küchenhoff, LMU München
6. Conclusion

• Prevalence estimation based on kappa is an instance where the idea of partial identification (econometrics) and systematic sensitivity analysis (biometrics) allows for reliable but still informative conclusions.

• Introducing successively additional assumptions would lead to smaller, but less credible regions.

• Optimize confidence interval (finite sample correction)!