Int J Public Health
DOI 10.1007/s00038-012-0439-9

HINTS & KINKS

Misclassification errors in prevalence estimation: Bayesian handling with care

Niko Speybroeck · Brecht Devleesschauwer · Lawrence Joseph · Dirk Berkvens

Received: 28 September 2012 / Accepted: 3 December 2012

This Hints and Kinks paper starts from the simple but well-known premise that "what gets measured, gets done", which we would like to extend into "what gets measured well, gets done well", and finally to "what does not get measured well could still get done well, if appropriate analytical methods are used".

Imagine assessing the prevalence of an infectious disease in a population, where the presence of disease is determined by a diagnostic test. For each tested individual, the diagnostic test result gives a "signal" that does not necessarily match its true infection status. It is well known that false positive and false negative results can arise when using diagnostic tests, for example producing a positive result in a non-case owing to a factor unrelated to the infection. On a population level, the prevalence as determined by the diagnostic test will thus only be an "apparent" prevalence, which will, to some extent, differ from the "true" prevalence. This problem of diagnostic test misclassification is a special case of information bias. We will describe the general problem and provide suggestions for adjusting for misclassification bias in practice.

The performance of a diagnostic test is typically parameterized by two quantities, the sensitivity and the specificity, each describing the capacity of the test to reflect the unknown "true" disease status. To define these two quantities, the following notation is introduced:

- infection status D: $D = 1$: infected, $D = 0$: not infected
- diagnostic test result Y: $Y = 1$: positive, $Y = 0$: negative

Combining each possible value of infection status and diagnostic test result leads to the cross table presented as Table 1.

In this table, the four possible combinations of infection status and diagnostic test result are denoted as follows:

- TP = True positive: truly infected individual with a positive test result
- TN = True negative: truly non-infected individual with a negative test result
- FP = False positive: truly non-infected individual with a positive test result
- FN = False negative: truly infected individual with a negative test result

Following the same notation, the "true" prevalence, denoted π, and the "apparent" prevalence, denoted p, are given by:

- True prevalence (π): $P(D = 1) = (TP + FN)/(TP + FP + FN + TN)$

N. Speybroeck (✉) · B. Devleesschauwer
Institute of Health and Society, Université catholique de Louvain, Brussels, Belgium
e-mail: Niko.Speybroeck@uclouvain.be

B. Devleesschauwer
e-mail: brecht.devleesschauwer@ugent.be

B. Devleesschauwer
Department of Virology, Parasitology and Immunology, Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium

L. Joseph
Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada
e-mail: lawrence.joseph@mcgill.ca

D. Berkvens
Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium
e-mail: dberkvens@itg.be

Published online: 20 December 2012

- Apparent prevalence (p): $P(Y = 1) = (TP + FP)/(TP + FP + FN + TN)$

The test sensitivity (SE) equals the probability that a truly infected individual will test positive, whereas the test specificity (SP) equals the probability that a truly non-infected individual will test negative. With "|" meaning conditional on (or "given"), SE and SP can be formally defined as:

- Test sensitivity (SE): $P(Y = 1 \mid D = 1) = TP/(TP + FN)$
- Test specificity (SP): $P(Y = 0 \mid D = 0) = TN/(FP + TN)$

By plugging in the above definitions for π, p, SE and SP, and setting the total population to 100 % (i.e., TP + FP + FN + TN = 1), Table 1 may be rewritten as Table 2.

From Table 2, it can be seen that the "apparent" prevalence (p) is related to the "true" prevalence (π) through the formula:

$$p = \pi \times SE + (1 - \pi) \times (1 - SP) \quad (1)$$

If the applied diagnostic test has perfect SE and SP, both equal to 100 %, Eq. (1) reduces to $p = \pi$. Indeed, as every infected individual will yield a positive test result and every non-infected individual a negative test result, the "apparent" prevalence (p) equals the "true" prevalence (π). In this paper, we call such a diagnostic test a gold standard test, although some researchers will use that term to denote the best reference test, even if its properties are imperfect.

Gold standard tests are rarely available, however, as most diagnostic tests are imperfect (SE ≠ 100 % and/or SP ≠ 100 %). To illustrate the effect of imperfect test characteristics on the "apparent" prevalence, let us suppose we wish to evaluate a population where the "true" infection prevalence equals 20 %. If the applied test has a less-than-perfect sensitivity, say 80 %, then p will equal 20 % × 80 % + (1−20 %) × (1−100 %) = 16 %. The "apparent" prevalence will thus underestimate the "true" prevalence. Otherwise, if the test has a less-than-perfect specificity, say 90 %, then p will equal 20 % × 100 % + (1−20 %) × (1−90 %) = 28 %. In this case, the "apparent" prevalence will overestimate the "true" prevalence. If both SE and SP are suboptimal, the "apparent" prevalence will be a result of both the false negative and false positive results. For example, combining both examples yields $p = 20\ \% \times 80\ \% + (1-20\ \%) \times (1-90\ \%) = 24\ \%$.

Table 1 Infection status by diagnostic test result cross table, in terms of TP, TN, FP, FN

Diagnostic test result	Infection status		
	D = 1	D = 0	
Y = 1	TP	FP	TP + FP
Y = 0	FN	TN	FN + TN
	TP + FN	FP + TN	TP + FP + FN + TN

D (infection status) = 1: infected, D = 0: not infected; Y (diagnostic test result) = 1: positive, Y = 0: negative

TP True positive, *TN* True negative, *FP* False positive, *FN* False negative

Table 2 Infection status by diagnostic test result cross table, in terms of π, p, SE and SP

Diagnostic test result	Infection status		
	D = 1	D = 0	
Y = 1	$\pi \times SE$	$(1-\pi) \times (1-SP)$	p
Y = 0	$\pi \times (1-SE)$	$(1-\pi) \times SP$	$1-p$
	π	$1-\pi$	1

π true (informed) prevalence, p apparent prevalence, *SE* test sensitivity, *SP* test specificity

The "true" infection prevalence, which we assumed to be 20 % in the above examples, will, of course, not be known in practical situations—else there would be no need to estimate the prevalence in the first place. In practice, the "true" prevalence is estimated from the "apparent" prevalence. Through algebraic manipulation of Eq. (1), we may obtain the Rogan–Gladen adjusted estimator of "true" prevalence (Rogan and Gladen 1978):

$$\pi = \frac{p + SP - 1}{SE + SP - 1} \quad (2)$$

Plugging in the "apparent" prevalence from our last example (24 %), together with the (presumed) known values of SE and SP, the "true" prevalence becomes:

$$\frac{24\% + 90\% - 1}{80\% + 90\% - 1} = 20\%$$

The Rogan–Gladen estimator, however, shows two important drawbacks. First, if the observed, "apparent" prevalence is lower than the probability of observing a false positive result (1−SP), the Rogan–Gladen estimate will become negative, which is of course implausible. As an example, suppose we obtained in our previous illustration an apparent prevalence of 5 % instead of 24 %. Plugging in this value into Eq. (2) would lead to:

$$\frac{5\% + 90\% - 1}{80\% + 90\% - 1} = -48\%$$

In each of the preceding examples, it was furthermore assumed that the values for SE and SP were fixed and known, for example provided by the test manufacturer. However, local factors, such as the presence of cross-reacting organisms, low infection pressure or the experience of the lab technicians can influence the SE or SP of a test. The characteristics of the test in the specific (field) conditions where the test will actually be used may

therefore differ from those obtained under the "ideal" laboratory conditions (with e.g., no cross-reactions due to other organisms). Therefore, context-specific test sensitivities and specificities will have to be used. These are, however, not known, and cannot be identified through the aforementioned Rogan–Gladen equation. Indeed, it is impossible to calculate three unknown quantities (π, SE and SP), from just one equation. Under certain conditions, using several diagnostic tests may partly solve this problem, but most often external information on the diagnostic test characteristics will be needed (Berkvens et al. 2006).

There are two statistical options for incorporating this external information in the analysis, originating from the frequentist or the Bayesian philosophy. In a frequentist analysis, the problem can in general only be circumvented by fixing the values of certain parameters. In the case at hand, this would mean setting the SE and SP equal to fixed, known values, as was done in the preceding illustrations. However, as highlighted in the foregoing paragraph, this is not optimal, as the SE and SP are context-specific and typically unknown, and should consequently be treated as random variables, not as fixed parameters. In the Bayesian philosophy, population parameters, including the SE and SP, are assumed to have intrinsic probability distributions, reflecting the uncertainty in their parameter values. The Bayesian approach further allows combining the observed field data (i.e., p) with any external (a priori) information on SE and SP within a single model. This external information can be historical information from experiments similar or related to the one under study or, in the absence of data, even beliefs of the investigators (e.g., expert opinions). The information on the test characteristics is thus not fixed anymore but expressed as a distribution or a range of values. Combining test results with a priori information on the test characteristics results in an a posteriori probability distribution of the prevalence.

If good prior information is available, the Bayesian method can thus flexibly account for parameter uncertainty in SE and SP while estimating π. Moreover, the Bayesian method makes it impossible to obtain negative values for π, since it is based on the transition from the true prevalence to apparent prevalence (i.e., Eq. (1)). This is further illustrated in Appendix 1, where the prevalence estimated as negative with the Rogan–Gladen estimator, is estimated as 0.48 %. However, it should also be clear that the results of the Bayesian method may strongly depend on the analysts summary of the available evidence to date about the sensitivity and specificity, expressed in the a priori distributions for SE and SP. Researchers should therefore be fully transparent about the distributions used in their models and about the methods used to derive these distributions. Estimating the posterior across the range of plausible values based on the evidence to date, may therefore help to assess the robustness of the estimates.

There are surprisingly few papers using the Bayesian approach for dealing with prevalence estimations in human medicine. The paper on malaria prevalence estimations in Peru, Vietnam and Cambodia (Speybroeck et al. 2011) was one of the first, if not the first, assessing the "true" malaria prevalence in a Bayesian manner. Other examples of using a Bayesian approach in estimating the prevalence are reported for strongyloidiasis, hepatitis E, and HIV infection (Joseph et al. 1995; Bouwknegt et al. 2008; Liu et al. 2011). Different ways to define expert opinions and to combine priors with the available data exist (Branscum et al. 2005; Berkvens et al. 2006; Engel et al. 2006).

The freeware programs WinBUGS (Lunn et al. 2000) and OpenBUGS (Lunn et al. 2009) are often used for Bayesian modelling. Appendix 1 provides an illustrative BUGS model for obtaining the "true" prevalence from an "apparent" prevalence based on individual samples. More code for assessing individual sample-based prevalences can be found in Berkvens et al. (2006), among others. Speybroeck et al. (2012) further provide code for obtaining the "true" prevalence from an "apparent" prevalence based on pooled samples, highlighting the flexibility and ease of extension of the Bayesian method.

In conclusion, if classification errors in the reference test are ignored, serious bias may be introduced in the assessment of the prevalence, although this is not always articulated as such. A Bayesian approach can be useful in this context, because it allows flexibly combining the available "prior" knowledge on diagnostic test characteristics with new data. Importantly, incorrect prior information can lead to unreliable posterior estimates. The use of several diagnostic tests may decrease the risk of errors, but most often the use of external information on the diagnostic test characteristics cannot be avoided. A reasonable option may therefore be to report results under different scenarios of test characteristics, and to be fully transparent about the applied test characteristics. In such a context, the term "true" prevalence may also be suboptimal, and we therefore prefer to call the obtained prevalence an "informed" prevalence. The use of such an "informed" estimation may avoid biased estimation of disease burden and may allow using surveillance systems more effectively when assessing for example the effects of interventions.

Appendix 1: Bayesian approach for estimating "true" prevalence from "apparent" prevalence

To introduce the Bayesian approach for estimating "true" prevalence (pi) from "apparent" prevalence (p), we build on the illustrations from the main text. In these examples, we had assumed a test sensitivity (SE) of 80 % and a test specificity (SP) of 90 %. We then estimated "true"

prevalence from an "apparent" prevalence of 24 and 5 %, using the Rogan–Gladen estimator (Eq. (2)). This resulted in estimates of, respectively, 20 and −48 %.

To parameterize the Bayesian model, we need information on the sample size (n), the number of positive test results (x), and an a priori probability distribution for "true" prevalence (pi). As a first illustration, we will continue to assume that SE and SP are known, fixed values (Model 1). This assumption will be relaxed in a further illustration (Model 2).

For simplicity, we will assume that our sample size was 500. An "apparent" prevalence of 24 % would thus have resulted from observing 120 positive test results. Likewise, an "apparent" prevalence of 5 % would have resulted from observing 25 positive results. As prior probability distribution for "true" prevalence, we will apply a uniform distribution ranging from 0 to 100 %. This is a common choice, as it expresses our belief before having observed any data, that the "true" prevalence can take any possible value and that each possible value is equally likely.

This information can now be used to establish the following Bayesian model:

Model 1 Bayesian estimation of true prevalence from apparent prevalence, based on fixed sensitivity and specificity; comments, denoted by the "#" symbol, translate the code in words

```
model {
  x ~ dbin(p, n)                      # 'x' results from a Binomial dist
                                      # with parameters 'p' and 'n'
  p <- pi * SE + (1 - pi) * (1 - SP)  # 'p' is related to 'pi'
                                      #  according to Equation (1)
  pi ~ dunif(0.00, 1.00)              # 'pi' follows a Uniform(0,1) dist
  SE <- 0.80                          # 'SE' equals 80%
  SP <- 0.90                          # 'SP' equals 90%
}
```

We fitted the model in WinBUGS using two chains, each containing 6,000 samples, of which the first 1,000 were discarded as "burn-in". Typical output is presented in Table 3. As expected, the Bayesian model succeeded in yielding only positive values, thus providing a useful "true" prevalence estimate for our second case.

The Bayesian approach makes it possible to flexibly account for uncertainty in the values for SE and SP. Instead of assuming fixed values, we assume in Bayesian model 2 that SE can take any possible value between 70 and 90 %, and that SP can take any possible value between 85 and 95 %. For the sake of comparability, these values were chosen so that their means would correspond to the fixed values applied before.

Table 3 "True" prevalence estimation results from Bayesian model 1

"Apparent" prevalence	Estimated "true" prevalence		
	Mean	2.5 %	97.5 %
0.24	0.2015	0.1493	0.2577
0.05	0.0048	0.0001	0.0173

Table 4 "True" prevalence estimation results from Bayesian model 2

"Apparent" prevalence	Estimated "true" prevalence		
	Mean	2.5 %	97.5 %
0.24	0.2008	0.1133	0.2927
0.05	0.0096	0.0003	0.0307

Model 2 Bayesian estimation of true prevalence from apparent prevalence, based on stochastic sensitivity and specificity; comments, denoted by the "#" symbol, translate the code in words

```
model {
  x ~ dbin(p, n)                      # 'x' results from a Binomial dist
                                      # with parameters 'p' and 'n'
  p <- pi * SE + (1 - pi) * (1 - SP)  # 'p' is related to 'pi'
                                      #  according to Equation (1)
  pi ~ dunif(0.00, 1.00)          # 'pi' follows a Uniform(0,1) dist
  SE ~ dunif(0.70, 0.90)          # 'SE' follows a Uniform(0.70, 0.90) dist
  SP ~ dunif(0.85, 0.95)          # 'SP' follows a Uniform(0.85, 0.95) dist
}
```

Again, the model was fitted in WinBUGS using two chains, each containing 1,000 "burn-in" samples and 5,000 retained samples. Typical (numerical and graphical) output is presented in Table 4. The results are similar to those obtained by Model 1, but show wider credibility intervals, owing to the additional uncertainty introduced by defining SE and SP as stochastic nodes.

References

Berkvens D, Speybroeck N, Praet N, Adel A, Lesaffre E (2006) Estimating disease prevalence in a Bayesian framework using probabilistic constraints. Epidemiology 17:145–153

Bouwknegt M, Engel B, Herremans MMPT, Widdowson MA, Worm HC, Koopmans MPG, Frankena K, De Roda Husman AM, De Jong MCM, Van Der Poel WHM (2008) Bayesian estimation of hepatitis E virus seroprevalence for populations with different exposure levels to swine in The Netherlands. Epidemiol Infect 136:567–576

Branscum AJ, Gardner IA, Johnson WO (2005) Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Prev Vet Med 68:145–163

Engel B, Swildens B, Stegeman A, Buist W, De Jong M (2006) Estimation of sensitivity and specificity of three conditionally dependent diagnostic tests in the absence of a gold standard. J Agric Biol Environ Stat 11:360–380

Joseph L, Gyorkos TW, Coupal L (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 141(3):263–272

Liu P, Xiao S, Shi ZX, Bi XX, Yang HT, Jin H (2011) Bayesian evaluation of the human immunodeficiency virus antibody screening strategy of duplicate enzyme-linked immunosorbent assay in Xuzhou blood center, China. Transfusion 51:793–798

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS—A Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325–337

Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique and future directions (with discussion). Stat Med 28:3049–3082

Rogan WJ, Gladen B (1978) Estimating prevalence from the results of a screening test. Am J Epidemiol 107:71–76

Speybroeck N, Praet N, Claes F, Van Hong N, Torres K, Mao S, Van den Eede P, Thi Thinh T, Gamboa D, Sochantha T, Thang ND, Coosemans M, Büscher P, D'Alessandro U, Berkvens D, Erhart A (2011) True versus apparent malaria infection prevalence: the contribution of a Bayesian approach. PLoS ONE 6(2):e16705

Speybroeck N, Williams CJ, Lafia KB, Devleesschauwer B, Berkvens D (2012) Estimating the prevalence of infections in vector populations using pools of samples. Med Vet Entomol 26:361–371