Epidemiol. Infect. (2010), **138**, 1674–1678. © Cambridge University Press 2010
doi:10.1017/S0950268810000385

Exact confidence limits for prevalence of a disease with an imperfect diagnostic test

J. REICZIGEL[1]*, J. FÖLDI[2] AND L. ÓZSVÁRI[3]

[1] *Szent István University, Faculty of Veterinary Science, Budapest, Hungary*
[2] *Intervet Hungary Ltd, Budapest, Hungary*
[3] *Ministry of Agriculture and Rural Development, Budapest, Hungary*

(*Accepted 21 January 2010; first published online 3 March 2010*)

SUMMARY

Estimation of prevalence of disease, including construction of confidence intervals, is essential in surveys for screening as well as in monitoring disease status. In most analyses of survey data it is implicitly assumed that the diagnostic test has a sensitivity and specificity of 100%. However, this assumption is invalid in most cases. Furthermore, asymptotic methods using the normal distribution as an approximation of the true sampling distribution may not preserve the desired nominal confidence level. Here we proposed exact two-sided confidence intervals for the prevalence of disease, taking into account sensitivity and specificity of the diagnostic test. We illustrated the advantage of the methods with results of an extensive simulation study and real-life examples.

Key words: Diagnostic test, exact confidence interval, prevalence of disease, sensitivity, specificity.

INTRODUCTION

Estimation of prevalence is a basic requirement in epidemiological studies. Authors usually accompany their observed sample prevalence with a 95% confidence interval (CI) for the population prevalence, to give an impression of the precision of the estimate [1, 2]. In most cases, however, the diagnostic test is imperfect, i.e. it has a sensitivity and/or specificity less than 100% [3, 4]. When the sensitivity and specificity of the test are known, the adjusted prevalence, also called true prevalence, is calculated according to the formula

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

where $prev_{\text{adj}}$, $prev_{\text{obs}}$, Se, and Sp denote adjusted prevalence, observed prevalence (also called apparent prevalence), sensitivity, and specificity, respectively [5, 6].

CIs are given in most papers only for the observed prevalence without adjustment for sensitivity and specificity. Usually the simplest asymptotic method, the Wald method is used [7, 8], but sometimes Wilson's score method [9] or the exact Clopper–Pearson method is applied [10–12].

CIs adjusted for sensitivity and specificity are rarely given, although Rogan & Gladen [5] described an appropriate modification of the asymptotic Wald method, which is also included in some textbooks [6], and implemented in some computer programs [13]. In spite of this, some authors calculate CIs naively,

* Author for correspondence: Dr J. Reiczigel, Szent István University, Faculty of Veterinary Science, Department of Biomathematics and Informatics, H-1078 Budapest, István u. 2, Hungary.
(Email: reiczigel.jeno@aotk.szie.hu)

adjusting the observed prevalence by the Rogan & Gladen formula, and then calculating a CI using the adjusted prevalence as if that were observed directly, i.e. without the necessary correction in the variance formula [14, 15]. This method results in incorrect CIs, whose actual coverage at a nominal level of 95% can be as low as 60%, even for large samples. The exact Clopper–Pearson CI has been applied with adjustment for sensitivity and specificity in some papers, although infrequently [16]. The problem regarding exact CIs for the adjusted prevalence is well illustrated by papers applying the exact Clopper–Pearson method for the observed and an asymptotic method for the adjusted prevalence [17].

Here we apply the approach proposed by Cameron & Baldock [18] to calculate exact two-sided CIs adjusted for sensitivity and specificity. As the Clopper–Pearson method is known to be too conservative for two-sided intervals [19, 20], we use Blaker's and Sterne's methods [20–22] providing shorter exact two-sided CIs.

METHODS

Both Blaker's and Sterne's CIs are derived by inverting the corresponding tests. Therefore we first defined how these tests can be adjusted for sensitivity and specificity. Both of them test for

$$H_0\colon p = p_{\text{hyp}} \quad \text{against} \quad H_1\colon p \neq p_{\text{hyp}},$$

where p and p_{hyp} denote the unknown true population prevalence and its hypothesized value, respectively.

Assuming H_0 and binomial sampling distribution, the probability that a random sample of size n contains k subjects with the disease ($k = 0, 1, \ldots, n$) is

$$\binom{n}{k} p_{\text{hyp}}^{k} (1 - p_{\text{hyp}})^{n-k}. \tag{2}$$

Given that the sample contains k subjects with the disease, and assuming that the diagnostic procedure has sensitivity Se and specificity Sp, we can calculate the probability that the number of subjects in the sample found to be positive is equal to m ($m = 0, 1, \ldots, n$). The formula for this is as follows:

$$\sum_{i=0}^{m} \binom{k}{i} Se^{i} (1 - Se)^{k-i} \binom{n-k}{m-i} \times (1 - Sp)^{m-i} Sp^{n-k-m+i}. \tag{3}$$

Combining equations (2) and (3), we observe that under H_0 the probability that the number of test positives in the sample is equal to m ($m = 0, 1, \ldots, n$) can be written as

$$P_{H_0}(m) = \sum_{k=0}^{n} \binom{n}{k} p_{\text{hyp}}^{k} (1 - p_{\text{hyp}})^{n-k} \times \left(\sum_{i=0}^{m} \binom{k}{i} Se^{i} (1 - Se)^{k-i} \times \binom{n-k}{m-i} (1 - Sp)^{m-i} Sp^{n-k-m+i} \right). \tag{4}$$

The principle of Sterne's method is that we order the sample space (i.e. the values $0, 1, \ldots, n$) according to their probabilities under H_0. From this it follows that the p value belonging to an observed number j of positives in the sample is

$$p_{j, H_0} = \sum_{P_{H_0}(i) \leqslant P_{H_0}(j)} P_{H_0}(i), \tag{5}$$

where $P_{H_0}(i)$ denotes the probability defined by equation (4). In Blaker's test, we order the sample space according to a so-called acceptability function

$$A_{H_0}(m) = \min \left(\sum_{i \leqslant m} P_{H_0}(i), \sum_{i \geqslant m} P_{H_0}(i) \right), \tag{6}$$

resulting in a p value of

$$p_{j, H_0} = \sum_{A_{H_0}(i) \leqslant A_{H_0}(j)} P_{H_0}(i). \tag{7}$$

Inversion of these tests results in exact two-sided CIs for prevalence. Test inversion means that observing j positives in the sample, the level $1 - \alpha$ CI runs from the smallest to the largest such p_{hyp} value, for which the test results in a p value P_{j, H_0} greater than α. Blaker's CI has the advantage that it is always contained in the Clopper–Pearson interval, whereas Sterne's CI, in spite of being slightly more narrow on average than Blaker's CI, may sometimes deliver intervals even wider than the Clopper–Pearson CI.

It should be noted that the proposed CIs do not have equal error probabilities in the two tails, thus one-sided intervals cannot be calculated from the two-sided ones in the usual way. For an exact one-sided CI the Clopper–Pearson method [23] should be used.

SIMULATION RESULTS

An extensive simulation study was performed to explore the coverage properties of the different methods and to compare the length of resulting 95% CIs. In the simulation we varied sensitivity and specificity

https://doi.org/10.1017/S0950268810000385 Published online by Cambridge University Press

Table 1. *Examples in which Blaker's CI proved to be narrower than Wilson's CI while its coverage probability was higher at the same time*

Sample size	True prevalence	Sensitivity	Specificity	95% Wilson's score interval		95% Blaker's interval	
				Coverage	Average length	Coverage	Average length
50	1%	95%	100%	91·6%	0·090	98·8%	0·085
100	1%	50%	100%	91·5%	0·089	98·9%	0·085
200	1%	50%	100%	92·1%	0·052	98·2%	0·051

Table 2. *Seroprevalence estimates of ovine paratuberculosis in sheep in selected regions in Portugal by Coelho et al. [12]. Seropositivity was determined using an ELISA test with sensitivity and specificity of 50% and 99·5%, respectively*

	Apparent prevalence	95% Clopper–Pearson CI	Estimated true prevalence	95% Clopper–Pearson CI	95% Sterne CI	95% Blaker CI
Boticas	2/78 = 2·6%	0·3–9·0%	4·2%	0–17·1%	0–16·8%	0–16·3%
Carrazeda e Vila Flor	4/130 = 3·1%	0·8–7·7%	5·2%	0·7–14·5%	1·1–14·3%	1·1–14·0%
Moimenta da Beira	8/78 = 10·3%	4·5–19·2%	19·7%	8·1–37·8%	8·7–37·6%	8·2–37·2%
Mogadouro	12/260 = 4·6%	2·4–7·9%	8·3%	3·9–15·0%	4·3–15·2%	4·1–15·0%
Vila Pouca	27/650 = 4·2%	2·8–6·0%	7·4%	4·6–11·1%	4·7–11·1%	4·6–11·0%

50%, 70%, 90%, 95%, 100%; true population prevalence 1%, 5%, 10%, 30%, 50%, 70%, and sample size 50, 100, 200, 500. From each combination of sensitivity, specificity, prevalence, and sample size, we generated 10 000 random samples and determined 95% CIs by the Wald, Wilson, Clopper–Pearson, Sterne, and Blaker methods. [Detailed results of the simulation study with respect to actual coverage (actual confidence level) and average length can be found on the website of the first author, http://www.univet.hu/users/jreiczig/prevalence-with-se-sp.html.] Simulation with 10 000 replications implies that the standard error of the obtained coverage probabilities is about 0·2%.

The actual coverage of the 95% Wald interval was often $<90\%$, in particular when prevalence was $<30\%$. In the worst cases, with low prevalence (1%), low sensitivity (50%), and high specificity (100%), the coverage was as low as 22·3% ($n=50$), 38·5% ($n=100$), 62·6% ($n=200$), and 85·9% ($n=500$). The actual coverage of the 95% Wilson interval was considerably better (worst case coverage was 90·8%, 91·5%, 92·1%, 93·9% for $n=50$, 100, 200, 500, respectively). Similarly to the Wald method, lowest coverage occurred in the case of low prevalence and sensitivity, combined with high specificity.

Exact methods produced in general longer intervals than the Wilson CI. In case of low prevalence, low sensitivity and high specificity the Sterne CI turned out to be even longer than the Clopper–Pearson interval. In these cases, Blaker's CI was the shortest, it was even shorter than the Wilson interval, in spite of the lower coverage of the latter (Table 1). If true prevalence was between 30% and 70%, the Sterne interval was the shortest among exact intervals.

EXAMPLE

Coelho *et al.* [12] conducted a survey to estimate the prevalence of ovine paratuberculosis in sheep flocks in the northeast of Portugal. Presence of antibodies against *Mycobacterium avium* subspecies *paratuberculosis* was investigated using a commercial enzyme-linked immunosorbent assay (ELISA) test. According to the manufacturer, the kit has sensitivity between 50% and 65% and specificity $>99·5\%$. These authors [12] present the seroprevalence values obtained by ELISA for each region, and also give exact Clopper–Pearson CIs for the apparent prevalence. However, they do not correct for test sensitivity and specificity. Table 2 illustrates that exactness of the CI does not prevent bias due to ignoring test imperfectness. The upper confidence limit adjusted for sensitivity and specificity turns out to be twice as high as the unadjusted one. Results without adjustment may lead to over-optimistic estimates of the infection status. Comparing the last three columns of Table 2,

https://doi.org/10.1017/S0950268810000385 Published online by Cambridge University Press

it can be seen that the Sterne interval is somewhat (by about 2·5%), and the Blaker interval slightly more (by about 3·5%) shorter than the Clopper–Pearson interval.

Further examples are available at the first author's website (see Simulation Results section).

DISCUSSION

Here we focused on CI construction, although the proposed methods can also be applied to two-tailed testing, leading to more powerful exact tests. It should be noted that the principles can also be applied to hypergeometric distribution as, i.e. to estimate prevalence in a finite population, adjusted for sensitivity and specificity.

It can be proved that transforming the exact lower and upper confidence limits obtained for apparent prevalence by the Rogan & Gladen formula (1) results in an exact CI for the true prevalence. This leads to an easy implementation of the method. Transformation of asymptotic confidence limits results of course in an asymptotic interval for the true prevalence with coverage comparable to that of the CI for the apparent prevalence. However, the naive method, in which the CI is calculated using the adjusted prevalence as if that value had been actually observed, turns out to be inappropriate. For example, simulating with sensitivity = 85%, specificity = 90%, and prevalence = 3% demonstrates that a 95% two-sided CI constructed in this way has an actual coverage probability of $<60\%$, regardless of whether an asymptotic or an exact CI calculation method is applied.

Computer programs (R functions as well as standalone programs for Microsoft Windows) for the described methods are available on the first author's website (see Simulation Results section), together with programs for sample size calculations to the proposed procedures. Note that the Sterne method is also worked out for the difference and ratio of two prevalences from independent samples, including tests and CIs [24]. A future task is to extend this to the case of imperfect tests allowing for different sensitivities and specificities.

How much more narrow the proposed CIs could be than the Clopper–Pearson interval depends on sensitivity, specificity, sample size and true prevalence. According to the simulation results, if true prevalence is between 30% and 70%, Sterne's CI is more narrow than Blaker's CI, whereas for $<20\%$ and $>80\%$ it is wider. Therefore, if we have prior information about the prevalence, we can choose between the methods. Note that here the expected true population prevalence is meant, not the observed sample prevalence. Choosing the method according to the observed sample prevalence will result in a CI worse than any one of the two methods.

Wilson's score interval, although asymptotic, performed much better than the Wald interval. It can be regarded as appropriate for sample sizes >500. The 92% worst case coverage for samples of $\leqslant 200$ cannot be compensated by its length. It was surprising that in some cases Blaker's CI was more narrow on average than Wilson's CI while its coverage was higher at the same time. A few illustrative examples are given in Table 1.

In extreme cases (prevalence = 0·01 or 0·02, sensitivity = 0·50, specificity = 1, $n = 50$), Sterne's CI was longer than the Clopper–Pearson interval. Blaker's CI has the advantage that this cannot occur. As its length never exceeds that of Sterne's CI, and that sometimes it is even narrower than Wilson's asymptotic interval, we propose Blaker's CI for general use.

It should be noted that the proposed methods assume that sensitivity and specificity are known exactly. These values are used in the procedure as fixed numbers known without any uncertainty. If they are estimated from an experiment of comparable size, then the uncertainty in sensitivity and specificity estimates should be accounted for in the CI construction, finally resulting in wider CIs. Rogan & Gladen [5] described how this can be done for the Wald interval, but currently it has not been developed for exact intervals.

CONCLUSION

Estimates of disease prevalence may be seriously biased if sensitivity and specificity of the diagnostic test are disregarded. In case of known sensitivity and specificity CIs are easy to adjust by applying the Rogan & Gladen transformation to the CI endpoints. Since asymptotic methods may not maintain the prescribed confidence level, we propose calculating exact CIs, in particular for sample sizes of $\leqslant 200$. Without any prior information on the true population prevalence we propose Blaker's method, as the resulting CI is always contained in the Clopper–Pearson CI. Note that if one-sided exact CIs are needed, the Clopper–Pearson interval is the only available option.

https://doi.org/10.1017/S0950268810000385 Published online by Cambridge University Press

DECLARATION OF INTEREST

None.

ACKNOWLEDGEMENTS

The authors thank the two anonymous referees for their comments that led to a considerable improvement of the paper.

REFERENCES

1. **Barrio G, *et al.*** Prevalence of HIV infection among young adult injecting and non-injecting heroin users in Spain in the era of harm reduction programmes: gender differences and other related factors. *Epidemiology and Infection* 2007; **135**: 592–603.
2. **Vyse AJ, Hesketh LM, Pebody RG**. The burden of infection with cytomegalovirus in England and Wales: how many women are infected in pregnancy? *Epidemiology and Infection* 2009; **137**: 526–533.
3. **Aguilar-Setién A, *et al.*** Dengue virus in Mexican bats. *Epidemiology and Infection* 2008; **136**: 1678–1683.
4. **Alonso-Padilla J, *et al.*** The continuous spread of West Nile virus (WNV): seroprevalence in asymptomatic horses. *Epidemiology and Infection* 2009; **137**: 1163–1168.
5. **Rogan WJ, Gladen B.** Estimating prevalence from the results of a screening test. *American Journal of Epidemiology* 1978; **107**: 71–76.
6. **Thrusfield MV.** *Veterinary Epidemiology*. New York: Wiley-Blackwell, 2007, pp. 610.
7. **Grossman Z, *et al.*** Absence of Kaposi sarcoma among Ethiopian immigrants to Israel despite high seroprevalence of human herpesvirus 8. *Mayo Clinic Proceedings* 2002; **77**: 905–909.
8. **Fernández-Limia O, *et al.*** Prevalence of *Candida albicans* and *Trichomonas vaginalis* in pregnant women in Havana City by an immunologic latex agglutination test. *Medscape General Medicine* 2004; **6**: 50.
9. **Garcia-Vazquez Z, *et al.*** Seroprevalence of *Neospora caninum* antibodies in beef cattle in three southern states of Mexico. *Tropical Animal Health and Production* 2009; **41**: 749–753.
10. **Moujaber T, *et al.*** The seroepidemiology of *Helicobacter pylori* infection in Australia. *International Journal of Infectious Diseases* 2008; **12**: 500–504.
11. **Claerebout E, *et al.*** Giardia and other intestinal parasites in different dog populations in Northern Belgium. *Veterinary Parasitology* 2009; **161**: 41–46.
12. **Coelho AC, *et al.*** Seroprevalence of ovine paratuberculosis infection in the Northeast of Portugal. *Small Ruminant Research* 2007; **71**: 298–303.
13. **Epi Tools.** Online pidemiological calculators, AusVet Animal Health Services, 2004. (http://epitools.ausvet.com.au). Accessed 10 July 2009.
14. **McCluskey BJ, *et al.*** A 3-year pilot study of sentinel dairy herds for vesicular stomatitis in El Salvador. *Preventive Veterinary Medicine* 2003; **58**: 199–210.
15. **O'Brien DJ, *et al.*** Estimating the true prevalence of *Mycobacterium bovis* in free-ranging elk in Michigan. *Journal of Wildlife Disease* 2008; **44**: 802–810.
16. **Thiry J, *et al.*** Serological evidence of caprine herpesvirus 1 infection in Mediterranean France. *Veterinary Microbiology* 2008; **128**: 261–268.
17. **Bartels CJ, *et al.*** Supranational comparison of *Neospora caninum* seroprevalences in cattle in Germany, The Netherlands, Spain and Sweden. *Veterinary Parasitology* 2006; **137**: 17–27.
18. **Cameron AR, Baldock FC.** A new probability formula for surveys to substantiate freedom from disease. *Preventive Veterinary Medicine* 1998; **34**: 1–17.
19. **Agresti A, Coull BA.** Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician* 1998; **52**: 119–126.
20. **Blaker H.** Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 2000; **28**: 783–798.
21. **Sterne TE.** Some remarks on confidence or fiducial limits. *Biometrika* 1954; **41**: 275–278.
22. **Reiczigel J.** Confidence intervals for the binomial parameter: some new considerations. *Statistics in Medicine* 2003; **22**: 611–621.
23. **Clopper CJ, Pearson ES.** The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**: 404–413.
24. **Reiczigel J, Abonyi-Tóth Z, Singer J.** An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio of proportions. *Computational Statistics and Data Analysis* 2008; **52**: 5046–5053.