

What can be learnt from nothing? – A statistical perspective

Olaf Gefeller, Annette B. Pfahlberg and Wolfgang Uter

Department of Medical Informatics, Biometry and Epidemiology, University of Erlangen-Nürnberg, 91054 Erlangen, Germany,

doi:10.1111/cod.12112

Summary

Background. No observed event is a special, but not uncommon, result in patch test studies. The interpretation of such findings depends critically on the sample size (n) of the investigation, and is statistically addressed by the use of confidence intervals (CIs).

Objective. To define the statistically correct method of calculating a CI with a confidence level of $1 - \alpha$, where α denotes the tolerated statistical error probability, for an observed prevalence of 0%.

Methods. A literature survey and evaluation of the statistical methods was conducted. The popular statistical software packages SPSSTM and SASTM were examined with regard to the methods implemented, and the results obtained, for estimating such CIs in this special case.

Results. The evaluation identified $[0; 1 - \alpha^{1/n}]$, which is well approximated by $[0; 3/n]$ for $\alpha = 0.05$, as an appropriate method to compute a CI with a confidence level of $1 - \alpha$. The resulting CI is an exact one, and more efficient than standard solutions. Popular statistical software such as SPSSTM and SASTM offers only various inefficient or even invalid procedures, but does not include this method.

Conclusions. It is easy to calculate a CI for an observed prevalence of 0% obtained in some studies. Such a CI facilitates the interpretation of such a finding, as it puts the observed zero result into adequate statistical perspective.

Key words: biostatistics; interval estimation; prevalence.

Often, a study in contact allergy (or other) research is conducted not only to describe the actual patients studied, but also to enable conclusions to be drawn regarding similar subjects (patients) in general. In such a case, statisticians view the study subjects as a random sample of an underlying target population of 'similar' subjects (issues such as selection or other types of bias arising during the recruitment of study subjects will not be addressed here). Hence, any prevalence of some condition, a risk quotient contrasting two prevalences, or whatever

measure of interest is calculated from the study data, has two meanings: (i) it is a description of the actual sample; and (ii) it is a so-called 'point estimate' of the true parameter of interest in the target population. However, regarding the latter aspect, repetitions of the study, with new samples, will probably give rise to slightly different results, that is, slightly different point estimates. It is therefore of interest to compute the range that those (hypothetically) repeated point estimates will most likely fall into. This range describes the interval in which the true parameter in the target population can be expected to lie. Based on the same study data used to arrive at a point estimate, a so-called 'interval estimate' can be calculated, which indicates the so-called 'confidence interval' (CI). A CI gives an idea of the precision of the estimated prevalence, risk quotient or other measure of interest with regard to the true parameter in the underlying target population.

Correspondence: Olaf Gefeller, Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Waldstr. 6, D-91054 Erlangen, Germany. Tel: +49 9131 8522750; Fax: +49 9131 8522721. E-mail: olaf.gefeller@imbe.med.uni-erlangen.de

Conflicts of interest: The authors have declared no conflicts.

Accepted for publication 6 March 2013

Guidelines exist for the statistical analysis of contact allergy data (1), which include some background for interval estimation in general, and for several applications of this concept in real studies in particular. The present article focuses on a special, but not uncommon, problem in estimating prevalences such as sensitization frequencies in contact allergy research. Whereas the general guidelines on interval estimation can be followed whenever the estimated prevalence lies somewhere above 0% but below 100%, the special cases at the boundaries, that is, 0% and 100%, deserve special attention and treatment. As these extreme cases have symmetrical characteristics, we concentrate on the 0% situation. In fact, 'no observed event' is not uncommon and leads to typical issues:

- A substance tested in a limited number of patients, for example (a few) hundred, may not have evoked any positive patch test reactions. Does this observation justify concluding that it is not an allergen when tested in patients with characteristics similar to those seen in the study?
- Certain adverse events (irritant reactions and active sensitization – not to mention even more serious adverse events such as death) have not been observed in a moderately sized clinical sample. Does this observation justify stating that there is 'no risk' of these events?

In the following, we will show how to deal with this special situation of interval estimation and offer a statistical solution to these questions. Considering the intrinsically complex nature of the problem of concluding something from nothing, we feel that our approach is as straightforward and easy as possible.

Statistical Methods

In general, CIs can be constructed for varying degrees of desired confidence, termed confidence level ($1 - \alpha$); most often α is set to 0.05, that is, 5%, and thus 95% CIs are commonly used in practical applications. Statistically speaking, a 95% CI for a prevalence estimated in a single study gives a range of prevalence values that has a probability of 0.95 of covering the true, albeit forever unknown, prevalence in replication studies when the process of sampling the units of observation from the target population and calculating the prevalence from the observed data is repeated again and again. The width of such intervals is influenced by the confidence level (the more confident, the wider), the study size (the larger the sample, the narrower the CI), and the magnitude of the

prevalence itself (in the middle of the admissible range, i.e. around 50%, CIs are wider than in the tails).

Statistical inference about the precision of estimated prevalences uses techniques of interval estimation for binomial proportions, as statisticians use the binomial distribution to describe this situation. For this purpose, a variety of formulae have been developed in the statistical literature (2). The different approaches can be classified into different groups: (i) exact versus approximate CIs, (ii) central versus non-central CIs.

The first distinction refers to underlying distributional assumptions when the CI is calculated, which is a quite complicated statistical topic. Exact CIs do not impose any distributional assumptions, as they are formulated with direct reference to the binomial distribution. Such calculations can, however, be numerically quite cumbersome, and do always lead to conservative intervals that have at least the designated confidence level. To simplify numerical calculations and to reduce the conservatism of the resulting CI, a number of approximate solutions have been suggested, based on the approximation of the binomial distribution to the normal distribution valid for large samples. The simple CI of the Wald method (3), which can be found in any introductory textbook on statistics [also in (1)], is one example of such an approximate CI that has been accompanied by several modified versions over the decades.

The second distinction refers to the strategy applied when constructing the CI around the point estimate derived from the data. Central CIs are constructed in a strict two-sided fashion around the point estimate, spending half of the tolerated error α above the point estimate and the other half below the point estimate. Non-central CIs do not impose this restriction; they only ensure that the tolerated error for falling outside the CI is bounded at α , thus guaranteeing the designated confidence level ($1 - \alpha$) inside the CI. The best-known exact CI, recommended by Clopper and Pearson (4), is also a central CI; however, non-central exact CIs have also been suggested, for example by Sterne (5).

The distinction between central and non-central CIs should not be confused with the better known distinction between symmetrical and non-symmetrical CIs. The property of symmetry focuses on the shape of the CI relative to the point estimate, that is, the prevalence observed in the study. Only when the distances from the upper and lower bounds of the CI, respectively, to the point estimate are identical is the CI called symmetrical; otherwise, it is a non-symmetrical CI. Symmetrical CIs can be central, but, in special cases, they can also be non-central. On the other hand, central CIs can also be both symmetrical and non-symmetrical, as the property

of centrality only addresses the strategy of constructing the CI, and not its result.

These general considerations on how to perform valid and efficient interval estimation for a binomial proportion need special refinement in our case concerned with a CI for a point estimate of zero. The long-lasting statistical debate and controversy about the 'best' method to choose in practical applications is of no relevance for our case, as here there is a clear-cut solution, yielding an appropriate CI.

Results

Whenever a CI for an estimated prevalence of zero with a confidence level of $1 - \alpha$ has to be calculated from study data, for example in contact allergy research, this can be done simply by use of the following formula:

$$CI = \left[0; 1 - \alpha^{\frac{1}{n}} \right]$$

with α as the tolerated error probability and n as sample size. The resulting interval is an exact CI; that is, the confidence level ($1 - \alpha$) is guaranteed for all sample sizes and does not rely on any distributional assumptions. Clearly, it is also a non-central and non-symmetrical CI, as the point estimate of the prevalence is identical to the lower bound of the CI. Owing to its property of non-centrality, this interval is always shorter – and thus more efficient – than the standard Clopper–Pearson exact CI.

Although the formula for the upper bound of the CI can be easily computed with a pocket calculator, even this formula can be approximated by a very simple term that allows the use of mental arithmetic without any technical equipment at all. Under the term 'rule of three', a simple approximation for the case of $\alpha = 0.05$ has been suggested (6):

$$\left[1 - \alpha^{\frac{1}{n}} \right] \cong \frac{3}{n}$$

The rationale and validity of this approximation for the situation most frequently encountered in practical applications have been examined in general (7). Its use gives an excellent approximation to the exact upper bound of the CI in sample sizes of ≥ 100 .

Figure 1a shows the upper bound of the CI computed for an estimated prevalence of 0% for sample sizes of 10–50 study subjects, and Fig. 1b shows this for sample sizes of 50 and 300 study subjects. Figure 1b also illustrates the general phenomenon that the gain in precision per further increase in sample size becomes smaller and smaller.

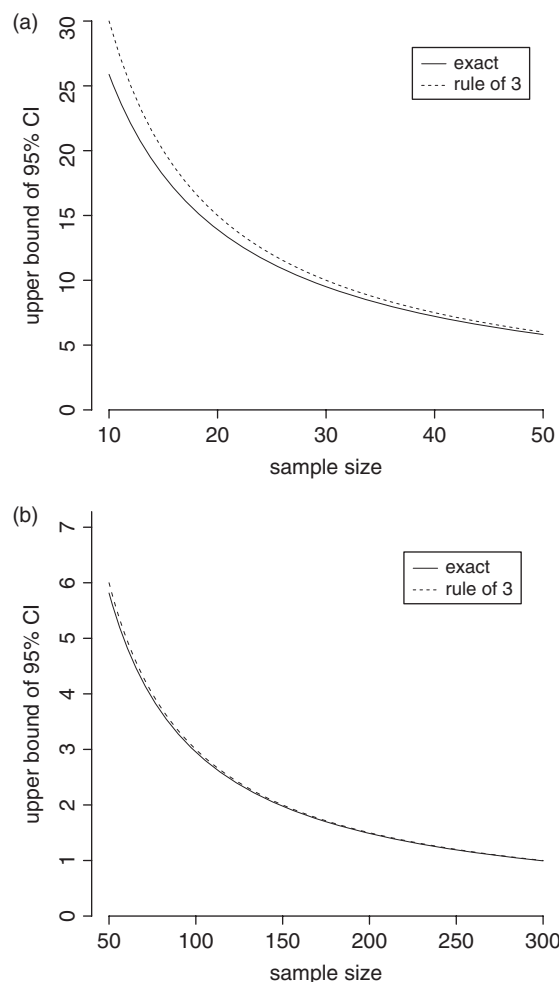


Fig. 1. Upper bounds of 95% confidence interval (CI) in percent for an observed prevalence of 0% with study size $n = 10$ – 50 (a) and $n = 50$ – 300 (b). The upper bound of the non-central exact 95% CI and its approximation by the 'rule of three' is shown – see text.

The implementation of interval estimation for binomial proportions in popular software packages for statistical analyses of contact allergy data, such as SPSSTM and SASTM, ignores the special boundary situation. Different CIs, such as the Clopper–Pearson exact CI (4) (SASTM and SPSSTM), the approximate CI according to the Wald method (3) (SASTM), the approximate CI according to the score method of Wilson (8) (SASTM) and its modified version of Agresti and Coull (9) (SASTM), a Bayesian CI based on Jeffreys non-informative priors (10) (SPSSTM), and a likelihood ratio test-based CI (11) (SPSSTM), are computed by the software packages, but none computes the non-central exact CI shown above.

As an illustration of the impact of using different formulae for estimating a CI from the same data in this

Table 1. Comparison of the results obtained with statistical software packages for calculation of an upper 95% confidence bound in per cent, to a prevalence of zero, in sample sizes of 50, 100, 200, 300, 500, and 1000, in relation to the method presented in this article (non-central exact) and its approximation (rule of three)

Method	n = 50	n = 100	n = 200	n = 300	n = 500	n = 1000
Wald	0	0	0	0	0	0
Wilson	7.13	3.70	1.88	1.26	0.76	0.38
Agresti–Coull	8.52	4.44	2.27	1.52	0.92	0.46
Jeffreys	4.88	2.47	1.25	0.83	0.50	0.25
Clopper–Pearson	7.11	3.62	1.83	1.22	0.74	0.37
Likelihood ratio test-based	3.77	1.90	0.96	0.64	0.39	0.19
Rule of three	6.00	3.00	1.50	1.00	0.60	0.30
Non-central exact	5.82	2.95	1.50	0.99	0.60	0.30

situation, Table 1 compares the six CI results computed by SPSSTM and SASTM with the non-central exact CI introduced above for five fixed sample sizes. The method most often referenced in textbooks, that is, the Wald method, yields a degenerated CI of 0–0 in this case, and is thus evidently invalid here. For smaller sample sizes, strong heterogeneity of the results is evident. Whereas the absolute magnitude of discrepancies between the methods becomes smaller for larger sample sizes, the relative differences remain high. For example, the width of the widest CI, interestingly resulting from the often advocated Agresti–Coull method, is > 50% larger than the exact CI suggested above.

Discussion

In the situation considered here, a CI captures the quantitative information that can be drawn from a study of limited size about the likelihood that an event of interest – not observed in the present study sample – may nevertheless occur in the target population to which the findings from the study are generalized. Our results show that the study size has to be quite large to confidently conclude that only a ‘negligible’ prevalence for observing the event of interest in the target population exists. In this context, the definition of ‘negligible’ will certainly depend on the severity of the event, or the graveness of its consequences.

We have presented a simple formula – and an even simpler approximation to it – for the calculation of a CI accompanying a zero prevalence for some event of interest; the application to construction of a lower bound of a CI for a 100% prevalence is symmetrical, and thus has not been dealt with further here.

The methodological discussion in the statistics community concerning interval estimation for a binomial proportion has centred on different topics that should not be confused with our special problem here. Non-central

exact CIs such as the one that we have suggested here do not constitute the ideal statistical solution to the general problem of computing CIs for a binomial proportion; they are, however, best suited for interval estimation at the boundaries, that is, 0% and 100%. This property has largely been neglected in the previous discussion in the statistics community; only Blyth (12) notes it in passing, echoed by Casella in his invited comments on a review paper of the topic (2). Probably because of this disregard of the boundary situation, popular statistical software packages do not give special consideration to the case where no (or all) events occur in the study, and provide only inefficient exact or invalid approximate CIs. The non-negligible variability of CIs resulting from the application of different methods in this situation shows that it is not only of theoretical interest to use a non-central CI as the only correct and efficient solution for the task of constructing a CI at the boundary.

This statistical finding has implications for the interpretation of results from studies in contact allergy. The notion of ‘no cases of sensitization seen, so the allergen is not a sensitizer’ has to be put into the perspective of the amount of empirical data available: in the case of a small study, the observed ‘negative result’ does not rule out the existence of a sensitization prevalence well above 0% among comparable patients. In addition to the statistical consideration outlined here, other aspects, such as selection of patients and appropriateness of patch testing, will, of course, also be of importance for the interpretation of such findings.

As one example from published data, van Oosten et al. noted no positive reaction to amyl cinnamal in 320 patients tested (13). The 95% CI for the sensitization prevalence ranged, however, up to 0.94% according to the above formula. In comparison, an analysis of data obtained by the Information Network of Departments of Dermatology (IVDK) in a similar study period found some positive reactions among 1214

patients consecutively tested with amyl cinnamal, namely a crude proportion of 0.25% (14), with, in this case, a 95% CI of 0.05–0.72%. Also, if a different subset of selectively tested patients included in the IVDK study is considered, with a higher crude prevalence of 0.72% positive reactions, the 95% CI is 0.41–0.90% for this subset. Evidently, the CIs of all three studies overlap broadly. Although the general interpretation of overlapping CIs needs special care (15), in this case it can be verified that the sensitization prevalences from the three analyses are not significantly different, or, in the words of statistical language, do not indicate that the three samples are drawn from target populations that differ with regard to the distribution of the outcome analysed.

Another example is a recent study on the appropriateness of patch test concentrations of selected fragrances (16), which stated that between $n = 103$ and $n = 114$ patients had been tested with certain patch test preparations, each 'without any evidence of active sensitization'. However, zero events in such a range of moderate sample sizes translate to an upper bound of the 95% CI of approximately 2.6–2.9%, according to the above formula. This implies that the observed negative study results regarding this adverse effect do not rule out, with the commonly accepted error of 5%,

that approximately 1 in 35 patients could be actively sensitized by the patch test preparations used.

Human experimental sensitization studies, such as those using human repeated insult patch tests (HRIPTs), are now considered to be unethical (17), although existing, historical data should not be ignored, for ethical reasons. Often, negative results, usually obtained in a few dozen to up to ~100 volunteers, are taken as evidence for a substance not being a contact sensitizer. As one example of numerous other studies, a HRIPT using the fragrance 4-methyl-3-decen-5-ol (CAS 81782-77-6), reviewed in (18), found no case of sensitization in 50 volunteers. The upper 95% confidence limit for this event rate is 5.8%. This implies that it cannot be ruled out (with 5% error) that approximately 1 in 18 healthy volunteers might be sensitized under the exposure conditions of the HRIPT.

Obviously, a reasonable balance between the resources devoted to a study and the precision achieved with a given sample size has to be found. However, the consequences of (necessarily) small study sizes in terms of precision need to be addressed in the reporting of study results.

With these examples alone, we hope to have convinced contact allergy researchers that, also for 'no event', the calculation – which is simple enough – of an appropriate CI, as described here, is one step further towards 'good epidemiological practice'.

References

- 1 Uter W, Schnuch A, Gefeller O. Guidelines for the descriptive presentation and statistical analysis of contact allergy data. *Contact Dermatitis* 2004; **51**: 47–56.
- 2 Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci* 2001; **16**: 101–133.
- 3 Wald A, Wolfowitz J. Confidence limits for continuous distribution functions. *Ann Math Statist* 1939; **10**: 105–118.
- 4 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**: 404–413.
- 5 Sterne TE. Some remarks on confidence or fiducial limits. *Biometrika* 1954; **41**: 275–278.
- 6 Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything alright? Interpreting zero numerators. *JAMA* 1983; **249**: 1743–1745.
- 7 Jovanovic BD, Levy PS. A look at the rule of three. *Am Stat* 1997; **51**: 137–139.
- 8 Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 1927; **22**: 209–212.
- 9 Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. *Am Stat* 1998; **52**: 119–126.
- 10 Rubin DB, Schenker N. Logit-based interval estimation for binomial data using the Jeffreys prior. In: *Sociological Methodology*, Clogg CC (ed): Washington, DC, American Sociological Association, 1987; pp. 131–144.
- 11 Anonymous. Nonparametric test algorithms. In: *IBM SPSS Statistics 20 Algorithms*, Anonymous (ed): Chicago, IL, IBM Corporation, 2011; pp. 644–673.
- 12 Blyth CR. Approximate binomial confidence limits. *J Am Stat Assoc* 1986; **81**: 843–855.
- 13 van Oosten EJ, Schuttelaar ML, Coenraads PJ. Clinical relevance of positive patch test reactions to the 26 EU-labelled fragrances. *Contact Dermatitis* 2009; **61**: 217–223.
- 14 Uter W, Geier J, Frosch PJ, Schnuch A. Contact allergy to fragrances: current patch test results (2005 to 2008) from the IVDK network. *Contact Dermatitis* 2010; **63**: 254–261.
- 15 Cumming G. Inference by eye: reading the overlap of independent confidence intervals. *Stat Med* 2009; **28**: 205–220.
- 16 Bruze M, Svedman C, Andersen KE et al. Patch test concentrations (doses in mg/cm²) for the 12 non-mix fragrance substances regulated by European legislation. *Contact Dermatitis* 2012; **66**: 131–136.
- 17 SCCP. Opinion concerning the predictive testing of potentially cutaneous sensitizing cosmetic ingredients or mixtures of ingredients adopted by the SCCNFP during the 11th plenary session of 17 February 2000, 2000.
- 18 McGinty D, Letizia CS, Api AM. Fragrance material review on 4-methyl-3-decen-5-ol. *Food Chem Toxicol* 2010; **48** (Suppl. 3): S93–96.