Biometrical Journal **47** (2005) 2, 237–247 DOI: 10.1002/bimj.200410104

Tutorial: Using Confidence Curves in Medical Research

Ralf Bender[*,1], **Gabriele Berg**[2], and **Hajo Zeeb**[2]

[1] Institute for Quality and Efficiency in Health Care, Cologne, Germany
[2] Department of Epidemiology and Medical Statistics, School of Public Health, University of Bielefeld, Germany

Received 20 February 2004, revised 17 November 2004, accepted 8 December 2004

Summary

Confidence intervals represent a routinely used standard method to document the uncertainty of estimated effects. In most cases, for the calculation of confidence intervals the conventional fixed 95% confidence level is used. Confidence curves represent a graphical illustration of confidence intervals for confidence levels varying between 0 and 100%. Although such graphs have been repeatedly proposed under different names during the last 40 years, confidence curves are rarely used in medical research. In this paper, we introduce confidence curves and present a short historical review. We draw attention to the different interpretation of one- and two-sided statistical inference. It is shown that these two options also have influence on the plotting of appropriate confidence curves. We illustrate the use of one- and two-sided confidence curves and explain their correct interpretation. In medical research more emphasis on the choice between the one- and two-sided approaches should be given. One- and two-sided confidence curves are useful complements to the conventional methods of presenting study results.

Key words: Confidence curves; Confidence intervals; Effect size; Graphical methods; Hypothesis testing; One-sided hypotheses; Statistical significance.

1 Introduction

Confidence intervals are widely used in medical research to document the uncertainty of estimated effects (Altman et al., 2000). In most applications, confidence intervals are calculated by using the conventional fixed 95% confidence level. By applying various confidence levels and representing the results graphically as confidence curves, more information can be extracted from the data. Although such graphs have been repeatedly proposed under different names during the last 40 years (Cox, 1958; Birnbaum, 1961; Kempthorne and Folks, 1971; Folks, 1981; Miettinen, 1985; Poole, 1987a, 1987b; Mau, 1988; Sullivan and Foster, 1990; Smith and Bates, 1992; Borenstein, 1994; Shakespeare et al., 2001), confidence curves are rarely used in the scientific literature. However, confidence curves are helpful tools for interpreting statistical results in medical research. As is the case with significance tests and confidence intervals, a choice between the one-sided and two-sided approach is required also for confidence curves to ensure adequate interpretation and conclusions. In this paper, we explain and discuss the use of one- and two-sided confidence curves in the analysis and reporting of medical studies.

2 One- and Two-sided Hypotheses

There is a controversy in the medical literature concerning the use of one- and two-sided significance tests (Dunnett and Gent, 1996). In the current practice of significance testing in clinical and epidemiological research two-sided P-values are presented in most cases (Bland and Altman, 1994). However,

[*] Corresponding author: e-mail: Ralf.Bender@iqwig.de, Phone: +49 221 35685-451, Fax: +49 221 35685-891

there are situations in which a one-sided approach is appropriate, for example in non-inferiority trials (ICH E9 Expert Working Group, 1999). As the one-sided approach requires a smaller sample size than the two-sided counterpart, the one-sided option should be used more frequently in medical research (Knottnerus and Bouter, 2001). Important prerequisites for the correct use of one-sided tests are firstly to state the corresponding one-sided hypothesis in advance, and secondly to give reasons for choosing a one-sided approach in the study protocol.

Let δ be the parameter of interest (e.g. difference of means or risk ratio) and δ_0 the chosen null value for the effect, a one-sided test problem is given by

$$H_0: \delta \leq \delta_0 \quad \text{vs.} \quad H_1: \delta > \delta_0 \quad \text{or} \quad H_0: \delta \geq \delta_0 \quad \text{vs.} \quad H_1: \delta < \delta_0 \tag{1}$$

whereas a two-sided formulation of a test problem can be described by

$$H_0: \delta = \delta_0 \quad \text{vs.} \quad H_1: \delta \neq \delta_0\,. \tag{2}$$

To illustrate the differences between two- and one-sided statistical inference, we consider the situation of the WHO melanoma study (Cascinelli et al., 1998). This international multicenter randomized trial was carried out by the WHO Melanoma Program from 1982 to 1989. A sample of 252 patients with a primary melanoma on the trunk with no evidence of regional node or distant metastases and Breslow thickness of 1.5 mm or more was included. Patients were randomized to receive either immediate node dissection or node dissection delayed until clinical diagnosis of regional node metastases. Of the 252 patients entered, 12 patients were excluded and 240 were eligible and evaluable for data analysis. As effect measure to describe the effect of immediate or delayed dissection of regional nodes on the survival of patients the hazard ratio (HR) was used. The HR was estimated by means of Cox regression analysis adjusting for sex, Breslow thickness, and age. A two-sided hypothesis was formulated with the null hypothesis given as: "There is no difference in survival between patients after immediate or delayed nodal dissection ($HR = 1$)." The corresponding alternative hypothesis was: "There is a difference in survival between the two groups ($HR \neq 1$)." Given the two-sided hypothesis, the application of two-sided significance tests and confidence intervals is required.

If a one-sided approach had been chosen, the focus would have been on only one direction of the effect, either survival benefit or survival detriment. In order to investigate a potential survival benefit as main focus of interest, the null hypothesis "There is no difference in survival between the groups or an increased risk of death for patients after immediate dissection ($HR \geq 1$)", should be tested against the corresponding alternative hypothesis "There is a decreased risk of death for patients with immediate dissection ($HR < 1$, i.e. survival benefit)." For this one-sided hypothesis, the use of one-sided significance tests and one-sided confidence limits would be appropriate.

The choice of a one-sided approach means that the other direction of the effect can not be investigated with the same data. If unexpectedly the data show the reverse direction of the effect, the confirmatory testing of this hypothesis requires the observation of new data. By accepting this limitation, the whole significance level α can be spent on one side of the parameter space leading to a higher power or, alternatively, to a smaller sample size (Knottnerus and Bouter, 2001).

3 One- and Two-sided Confidence Intervals

Interestingly, while the use of one-sided tests already is a rarity, one-sided confidence limits are even less frequently used (exceptions are non-inferiority trials). If confidence intervals are presented, they are almost always two-sided, even in cases in which one-sided tests are used. For example, in a case-control study investigating the association between low-dose oral contraceptives and myocardial infarction, two-sided 95% confidence intervals were presented, although sample size calculations were based upon a one-sided significance test (Sidney et al., 1996). When confidence intervals became a standard method in the medical literature for the presentation of study results about 15 years ago (Gardner and Altman, 1986; Simon, 1986; Morgan, 1989), the fact that there is the choice between two-sided confidence intervals and one-sided confidence limits seems to have been ignored.

Returning to the WHO study (Cascinelli et al., 1998), we illustrate the difference between two- and one-sided confidence intervals. Since the study hypothesis was two-sided, the authors reported a two-sided confidence interval for the HR of mortality in the immediate dissection group compared to the delayed dissection group. As before, let δ be the parameter of interest, d an approximately normally distributed estimate of δ, $\mathrm{SE}(d)$ the standard error of d, and z_p the p-quantile of the standard normal distribution. With this notation, the usual two-sided Wald confidence interval at level $1-\alpha$ is given by

$$\mathrm{LL}(\delta) = d - z_{1-\alpha/2} \times \mathrm{SE}(d), \qquad \mathrm{UL}(\delta) = d + z_{1-\alpha/2} \times \mathrm{SE}(d)\,. \tag{3}$$

Frequently, the parameter of interest is a function of a regression coefficient from an appropriate regression model. In the Cox regression model, the hazard ratio is given by $\mathrm{HR} = \exp(\beta)$. A confidence interval for HR can be obtained by transforming the confidence limits of β. Let b be an estimate of β with standard error $\mathrm{SE}(b)$, a two-sided confidence interval at level $1-\alpha$ for HR is given by

$$\mathrm{LL}(\mathrm{HR}) = \exp\left[b - z_{1-\alpha/2} \times \mathrm{SE}(b)\right], \qquad \mathrm{UL}(\mathrm{HR}) = \exp\left[b + z_{1-\alpha/2} \times \mathrm{SE}(b)\right]. \tag{4}$$

In the WHO study, the 95% confidence interval for the estimated $\mathrm{HR} = 0.72$ had a lower limit (LL) of $\mathrm{LL} = 0.49$ and an upper limit (UL) of $\mathrm{UL} = 1.04$ and was not significant ($\mathrm{P} = 0.07$) at the conventional 5% level, because the interval [0.49, 1.04] included the zero effect of $\mathrm{HR} = 1$ (Cascinelli et al., 1998). The confidence interval contains more information than the P-value, because it illustrates the range of possible treatment effects, which are compatible with the observed data at confidence level 95% in both directions (survival benefit and survival detriment). Strictly speaking, the confidence level is the probability that the confidence interval covers the true parameter of interest before the data are observed. After the calculation of a confidence interval from observed data no probabilty statements can be made in a strict mathematical sense. We express this by using the terminolgy that there is a confidence of 95% that the true HR lies within the interval between 0.49 to 1.04.

In the two-sided approach, the $(1-\alpha/2)$-quantile of the standard normal distribution is entered into the calculations, whereas the one-sided approach uses the $(1-\alpha)$-quantile leading to different values for the confidence limits. The one-sided upper Wald confidence limit at level $1-\alpha$ for a parameter δ of interest with approximately normally distributed estimate d is given by

$$\mathrm{UL}(\delta) = d + z_{1-\alpha} \times \mathrm{SE}(d)\,. \tag{5}$$

Similar to formula (4) the one-sided upper confidence limit of HR at level α is given by

$$\mathrm{UL}(\mathrm{HR}) = \exp\left[b + z_{1-\alpha} \times \mathrm{SE}(b)\right]. \tag{6}$$

For the one-sided hypothesis in the WHO study as outlined above the one-sided 95% limit $\mathrm{UL} = 0.98$ is obtained. Thus, in contrast to the two-sided approach, for the one-sided hypothesis a significant survival benefit is obtained, because the UL is smaller than 1. Moreover, the UL gives the information that – with confidence 95% – the smallest possible benefit is given by $\mathrm{HR} = 0.98$.

4 One- and Two-sided Confidence Curves

The values of confidence intervals depend on the arbitrary choice of the confidence level. The common choice for the confidence level in medical research is $1-\alpha = 95\%$. By varying the confidence level between 0 and 100% and plotting these values on the y-axis together with the corresponding confidence limits on the x-axis, a graph called *confidence curve* is obtained. The use of two-sided confidence intervals leads to two-sided confidence curves, whereas one-sided confidence limits for varying levels can be displayed by means of a one-sided confidence curve.

To produce such graphs, the formulas of the corresponding confidence limits have to be solved for the confidence level $1-\alpha$, so that the confidence level is formally presented as a function of the considered effect measure. Let $\Phi(\cdot)$ be the distribution function of the standard normal distribution. The two functions given in formula (3) can be solved separately for $1-\alpha$. The resulting two func-

tions have two different domains $]\infty, d]$ and $[d, \infty[$ but the same range. Thus, the confidence level $1 - \alpha$ can be presented as a function of the effect measure by means of the equation

$$1 - \alpha = 2\Phi\left(\frac{|d - \tilde{\delta}|}{\mathrm{SE}(d)}\right) - 1 \tag{7}$$

where $\tilde{\delta}$ denotes an arbitrary point in the parameter space of δ. Eq. (7) defines the two-sided confidence curve for the effect measure δ. Setting $\mathrm{UL}(\delta) = \tilde{\delta}$ in Eq. (5) and solving for $1 - \alpha$ leads to

$$1 - \alpha = \Phi\left(\frac{\tilde{\delta} - d}{\mathrm{SE}(d)}\right) \tag{8}$$

which gives the one-sided confidence curve corresponding to UL.

Other methods for confidence interval calculation lead to other equations for the corresponding confidence curves. If the effect measure of interest is a function of a parameter β for which an approximately normally distributed estimate b is available, an appropriate transformation is required. For example, if the relative survival effect $\mathrm{RSE} = (1 - \mathrm{HR}) \times 100$ (in %) with $\mathrm{HR} = \exp(\beta)$ is chosen as effect measure, we get the equation

$$1 - \alpha = 2\Phi\left(\frac{\left|\log\left(1 - \frac{\tilde{\mathrm{RSE}}}{100}\right) - b\right|}{\mathrm{SE}(b)}\right) - 1 \tag{9}$$

for the two-sided and

$$1 - \alpha = \Phi\left(\frac{\log\left(1 - \frac{\tilde{\mathrm{RSE}}}{100}\right) - b}{\mathrm{SE}(b)}\right) \tag{10}$$

for the one-sided confidence curve.

The two-sided confidence curve for the data of the WHO melanoma study (Cascinelli et al., 1998) regarding the effect measure RSE is shown in Figure 1. The relative survival effect of immediate nodal dissection in patients with truncal melanoma is shown on the x-axis. For the confidence level of 0% the estimated relative survival effect of $\mathrm{RSE} = 28\%$ is obtained. Increasing the confidence level leads to a wider confidence interval but also increases the confidence that the interval covers the true relative survival effect. The confidence interval for the conventional 95% confidence level is marked by the significance line and includes a relative survival effect ranging from -4% to $+51\%$, demonstrating a non-significant result in the two-sided setting as described above. However, it can be argued that immediate nodal dissection is beneficial, because up to a confidence level of 93% (corresponding to the two-tailed $\mathrm{P} = 0.07$) all values lie in the right half of the diagram representing survival benefit.

The one-sided confidence curve to investigate survival benefit of nodal dissection is shown in Figure 2. At confidence level 50% the curve shows the point estimate of the relative survival effect ($\mathrm{RSE} = 28\%$). With confidence 89% (93%) the real survival benefit is at least 10% (5%). In the one-sided setting, the data reveal a significant result at level $\alpha = 0.05$ because the confidence curve crosses the corresponding significance line. With confidence 95% the real relative survival benefit is at least 2%. The statement that there is any survival benefit of immediate nodal dissection in patients with truncal melanoma can be made with a confidence level of 96.5%.

For the same data a confidence curve was also presented by Shakespeare et al. (2001, Figure 3, p. 1351). However, in this one-sided confidence curve incorrect significance lines were used. The conventional significance level of $\alpha = 0.05$ corresponds to a confidence level of $1 - 0.05 = 95\%$, not

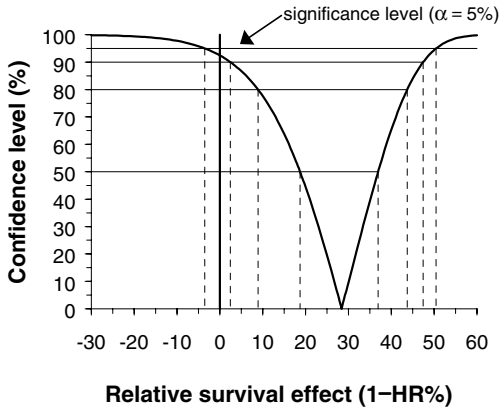

Figure 1 Two-sided confidence curve for the relative survival effect of nodal dissection in the WHO melanoma study (Cascinelli et al., 1998).

to 97.5%, as noted incorrectly by Shakespeare et al. (2001). Therefore, the significance line should be placed at 95% leading to a significant effect of nodal dissection in a one-sided setting (Figure 2). Whether a one-sided confidence curve is appropriate depends on the investigated hypothesis. According to the two-sided hypothesis in the WHO study, the two-sided confidence curve shown in Figure 1 is appropriate here.

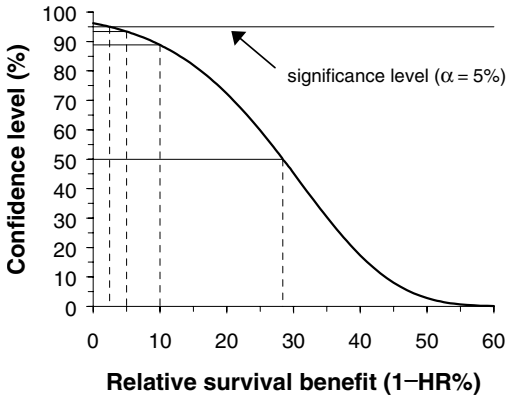

Figure 2 One-sided confidence curve for the relative survival benefit of nodal dissection in the WHO melanoma study (Cascinelli et al., 1998).

5 Application of Confidence Curves

5.1 General comments

The presentation of study results by means of confidence curves is useful whenever confidence intervals are appropriate and the consideration of various confidence levels is desirable. The proposed

method to create the corresponding graphs by presenting the confidence level as a function of the considered effect measure is applicable whenever closed formulas to calculate confidence intervals exist, which can be solved for $1 - \alpha$. This is usually the case if the considered confidence interval formula is based upon large sample approximations, such as the Wald method. However, there are also methods to calculate confidence intervals, which require iterative algorithms for solution. For example, Lui and Lin (2003) compared several asymptotic interval estimators for the odds ratio in a 2×2 table. For the Woolf and the Gart method (see Lui and Lin, 2003) it is possible to present the confidence level as function of the odds ratio. However, the application of the Cornfield method (with or without continuity correction) requires the iterative solution of two equations (see Lui and Lin, 2003). In such cases, it is still possible to create confidence curves. One has to apply the iterative algorithm to a large number of confidence levels varying between 0 and 1 and to create the confidence curve by interpolating the calculated data points. This procedure is straightforward, but cumbersome.

The formulas to create confidence curves proposed in this paper are based on the normal distribution. Thus, in most cases, they are adequate only for large samples. In small samples, the application of exact methods may be required. In the case of continuous data, the formulas have to be modified by using corresponding appropriate distributions, such as the t or F distribution. In the case of discrete data there are no definite confidence intervals that are clearly optimal with coverage probability precisely equal to the nominal level. Exact confidence curves for categorical data based upon uniformly most powerful unbiased tests have been proposed by Scherb and Brüske-Hohlfeld (1993). In this approach no single curve is produced. These exact confidence curves consist of two bounds based upon the ranges of lower and upper randomized confidence limits, which take the inherent uncertainty caused by discreteness into account. Blaker (2000) proposed a general method to construct exact confidence curves for discrete distributions on the basis of inverting exact tests with specific acceptance regions. The latter method is inherently two-sided and can therefore be used only for the construction of two-sided confidence curves. A discussion of the additional characteristics of these exact confidence curves is beyond the scope of this paper. The reader is referred to Scherb and Brüske-Hohlfeld (1993) and Blaker (2000) for a comprehensive discussion of exact confidence curves for parameters of discrete distributions.

5.2 Example

We present another example to illustrate the usefulness of confidence curves for the presentation of study results. Lui (2000) discussed the interval estimation of the difference between the marginal probability of a primary infection and the conditional probability of the secondary infection, given the primary infection, to assess the effect of the primary infection on the risk of developing the secondary infection. This situation leads to dependent samples, so that the usual methods for independent samples are not applicable. Lui (2000) developed three asymptotic interval estimators based upon the Wald test, the likelihood ratio test, and Fieller's theorem. Here, we consider only the Wald-test based confidence interval. Let π_{ij} $(i,j = 1,2)$ denote the probability of the corresponding cells in the following 2×2 table.

Table 1 2×2 table in studies investigating the risk of a secondary infection, given the primary infection.

		Secondary infection		
		Yes	No	
Primary infection	Yes	π_{11}	π_{12}	$\pi_{1\bullet}$
	No	–	π_{22}	π_{22}

In terms of the π_{ij}, the difference between the marginal probability of a primary infection and the conditional probability of the secondary infection, given the primary infection, can be described by

$$\delta = \pi_{1\bullet} - \frac{\pi_{11}}{\pi_{1\bullet}} \tag{11}$$

Let $n_{ij}\,(i,j = 1,2)$ denote the numbers of observed subjects in the corresponding cells of the 2×2 table above. Then the difference of interest δ can be estimated by means of

$$d = p_{1\bullet} - \frac{p_{11}}{p_{1\bullet}} \tag{12}$$

where $p_{11} = n_{11}/n$ and $p_{1\bullet} = (n_{11} + n_{12})/n$. The two-sided Wald confidence interval for δ at level $1 - \alpha$ is given by (3), where the standard error of d is

$$\mathrm{SE}(d) = \sqrt{\frac{1}{n}\left(\frac{p_{11}p_{12}}{p_{1\bullet}^3} + p_{1\bullet}(1 - p_{1\bullet})\right)} \tag{13}$$

(Lui, 2000). With this notation we can directly apply formula (7) to create a two-sided confidence curve for δ.

We consider the data of a study investigating pneumonia infection in 156 calves (Lui, 2000). The observed numbers of calves are $n_{11} = 30$, $n_{12} = 63$, and $n_{22} = 63$ leading to the estimate $d = 0.274$ with $\mathrm{SE}(d) = 0.0624$. From the 95% Wald confidence interval [0.151, 0.396] Lui (2000) concluded that the primary infection of pneumonia should stimulate a natural immunity to reduce the risk of a secondary infection because $\mathrm{LL} = 0.151 > 0$. By using formula (7) we produced a two-sided confidence curve for δ (Figure 3). In addition to the conclusion above we can assess the possible amount of immunity in terms of δ for various confidence levels. In particular, we can conclude that it seems to be almost certain (confidence level larger than 99.9%) that the interval [0.05, 0.50] contains the true value of δ.

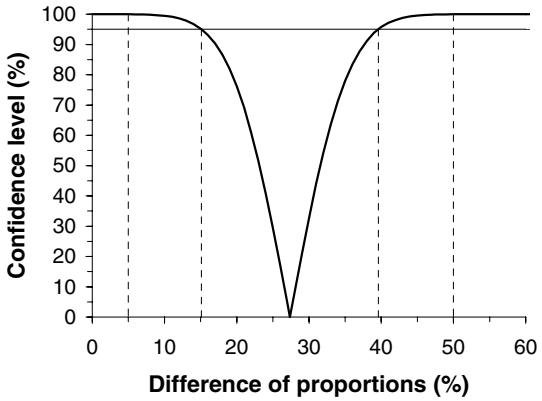

Figure 3 Two-sided confidence curve based upon Wald's statistic for the difference between the marginal probability of the primary infection with pneumonia and the conditional probability of the secondary infection, given the primary infection, estimated from 156 calves born in Florida (Lui, 2000).

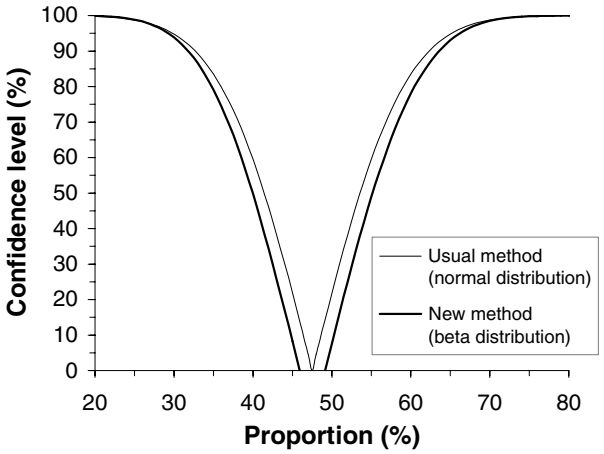

Figure 4 Two-sided confidence curves for the comparison of the usual procedure based upon the normal distribution and the new method proposed by Chen and Tipping (2002) based upon the beta distribution to calculate confidence intervals for a proportion with over-dispersion.

5.3 Comparison of confidence interval methods

Another useful feature of confidence curves is the possibility to compare different methods of confidence interval calculation over the whole range of confidence levels. For example, Chen and Tipping (2002) proposed a new procedure based upon the beta distribution to perform interval estimation of a proportion with over-dispersion and compared the new method with the regular Clopper-Pearson confidence interval ignoring over-dispersion, the exact confidence interval and the usual asymptotic Wald-type procedure. Here, we consider only the new method based upon the beta distribution and the Wald method for over-dispersed binary data. We used the simplified setting of a cluster randomized trial with $K = 20$ clusters and two binary observations in each cluster. We performed one of the simulations described by Chen and Tipping (2002), namely the situation of a marginal event rate of $\pi = 0.5$ and a correlation of $\rho = 0.3$ between the observations within one cluster.

Let n be the observed total number of events and $N = 2K$ the total number of observations. An estimate of π is given by $p = n/N$ with standard error

$$\mathrm{SE}(p) = \sqrt{\frac{1}{N}(1+\rho)\,p(1-p)} \tag{14}$$

(Chen and Tipping, 2002). The usual two-sided Wald confidence interval for π at level $1-\alpha$ taking over-dispersion into account is given by

$$\mathrm{LL}(\pi) = p - z_{1-\alpha/2} \times \mathrm{SE}(p), \qquad \mathrm{UL}(\pi) = p + z_{1-\alpha/2} \times \mathrm{SE}(p)\,. \tag{15}$$

Chen and Tipping (2002) proposed the new confidence limits

$$\mathrm{LL}(\pi) = b^{(1)}_{\alpha/2}\,, \qquad \mathrm{UL}(\pi) = b^{(2)}_{1-\alpha/2} \tag{16}$$

where $b^{(1)}_{\alpha/2}$ is the $\alpha/2$ quantile of the $\mathrm{Beta}\left(\dfrac{n}{1+\rho}, \dfrac{N-n}{1+\rho}+1\right)$ distribution and $b^{(2)}_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $\mathrm{Beta}\left(\dfrac{n}{1+\rho}+1, \dfrac{N-n}{1+\rho}\right)$ distribution. The confidence curve based upon the usual Wald

interval can be created by means of the formulas presented in chapter 4. The confidence curve for the new method based upon the beta distribution can be produced by means of the equation

$$1-\alpha = \begin{cases} 1-2B^{(1)}(\tilde{\pi}) & \text{if } B^{(1)}(\tilde{\pi}) \leq 0.5 \\ 2B^{(2)}(\tilde{\pi})-1 & \text{if } B^{(2)}(\tilde{\pi}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where $B^{(1)}(\cdot)$ and $B^{(2)}(\cdot)$ are the distribution functions of the beta distributions used in (16).

In an arbitrary simulation run of the data situation described above we observed $n = 19$ events leading to $p = 19/40 = 0.475$ with $\mathrm{SE}(p) = 0.09003$. Chen and Tipping (2002) compared the two methods only for the conventional 95% confidence level. They concluded that the actual coverage rates of the usual Wald procedure are generally below the target rate of 95%, while the proposed new procedure tends to have slighly higher coverage rates than 95%. With the two confidence curves shown in Figure 4 it is possible to compare the two methods over the whole range of possible confidence levels. It can be concluded that – in the considered data situation – both methods lead to similar limits for confidence levels of 95% and above, but that the two methods differ for lower confidence levels.

5.4 Medical decision making

In most applications, only the two-sided version of confidence curves was used. Exceptions are the graphs proposed by Kempthorne and Folks (1971), who considered one-sided as well as two-sided *families of consonance intervals*, the *confidence distribution function* of Mau (1988), which was derived from one-sided significance tests and the graphs proposed by Shakespeare et al. (2001), who constructed one-sided confidence curves although the considered hypothesis was two-sided. In the latter application, due to confusion about one- and two-sided statistical inference, incorrect significance lines were drawn in the graphs. The adequate application of confidence curves for medical decision making requires some prospective choices and statements. Firstly, it should be stated in advance whether the main study hypothesis is one- or two-sided and reasons for this decision should be given. Secondly, the significance level should be stated. There is no need to insist in general on the conventional 5% level. For example, in regulatory settings for clinical trials it is recommended to use a significance level of $\alpha = 2.5\%$ if one-sided tests are applied (ICH E9 Expert Working Group, 1999). These decisions determine whether a one- or two-sided confidence curve is required and at which level the significance line has to be placed.

In clinical and epidemiological trials it is common that beside the main study hypothesis there are a number of subsidiary hypotheses, which are investigated and tested in an exploratory manner. If the whole type 1 error probability is spent for the main study hypothesis, no part of the significance level is left for these subsidiary hypotheses. Nevertheless, for these subsidiary hypotheses confidence curves can be drawn. However, we recommend to plot these confidence curves without significance line and to label the curves clearly as exploratory result to distinguish them from the confirmatory analyses. If there is no single main study hypothesis and the whole type 1 error probability is distributed on several hypotheses (Moyé, 1998, 2000), the corresponding significance lines should be given in the graphs with a statement about this prospective alpha allocation. This procedure ensures that the resulting confidence curves can be used as a tool for an adequate medical decision making.

6 Discussion

For both one- and two-sided problems, it is a good idea to calculate confidence intervals for various levels instead of using only the conventional fixed 95% level (Cox, 1958). If a large number of different levels are used, the results should be presented in an appropriate graph. Birnbaum (1961) used the term "*confidence curves*" for such graphs. Kempthorne and Folks (1971) defined consonance intervals by inversion of significance tests and proposed graphs called "*family of consonance intervals*" for

both two- and one-sided significance tests. Folks (1981) noted that the calculation of consonance and confidence intervals leads to the same result, but that their interpretation is quite different. Using simultaneously the consonance as well as the confidence interpretation, Miettinen (1985) called this graph "*p-value function*". This term was adopted by several authors in the epidemiological literature (Poole, 1987a, 1987b; Foster and Sullivan, 1987; Smith and Bates, 1992). The problem with this term is the application of two different concepts in one graph. Thus, Sullivan and Foster (1990) recommended to use the term "*confidence interval function*". This term was also used by Borenstein (1994) in his introduction to confidence intervals. Mau (1988) applied a version of one-sided confidence curves to assess clinical equivalence. Referring to Cox (1958), Mau (1988) called this graph "*confidence distribution function*" due to similarities with a theoretical distribution function. The original term "*confidence curve*" was used by Scherb and Brüske-Hohlfeld (1993) and Blaker (2000), who developed exact methods to construct confidence curves for discrete data. Finally, Shakespeare et al. (2001) used the term "*clinical significance curve*" for the one-sided version of this graph.

Similar terms have also been proposed for other graphical procedures. Schweder and Spjotvoll (1982) proposed a graphical procedure, called "*P-value plot*" to evaluate multiple significance tests simultaneously. Hung et al. (1997) and Hung and O'Neill (2003) work with the "*P-value distribution function*". They interpret the P-value as a statistic, i.e. a function of the observed data, so that the theoretic cumulative distribution function of the P-value under the assumption that the alternative hypothesis is true can be derived. Both graphical procedures, the P-value plot and the P-value distribution function do not represent applications of confidence curves and are not considered here.

In the various applications of confidence curves different values and different ways of scaling the *y*-axis have been used. Possible choices are α ranging from 0 to 1 or from 1 to 0, $1-\alpha$ ranging from 0% to 100% or from 100% to 0%, and $\alpha/2$ (frequently called one-tailed p-value) ranging from 0 to 0.5 or from 0.5 to 0. By using the confidence interval interpretation we find it logical to use the confidence level $1-\alpha$ for the *y*-axis ranging from 0% to 100% as shown in Figures 1 to 4.

Which of the six proposed terms should be used for graphs of confidence intervals with varying confidence levels? Firstly, it should be recognized that the consonance interpretation has not found its way into medical research. If the confidence interpretation is applied, the terms *confidence curve*, *confidence interval function* and *confidence distribution function* are appropriate, whereas the term *p-value function* is misleading, because the confidence level is plotted on the *y*-axis and not the p-value. Secondly, the term *clinical significance curve* does not describe the real meaning of the graph. Whether the observed effect has clinical relevance can be hopefully derived by interpreting the graph but the content of the graph is statistical and not clinical by nature. In summary, we recommend to use the term *confidence curve* originally proposed by Birnbaum (1961).

After the development of theory of confidence sets by Neyman in 1930 (Folks, 1981), it took about 50 years until this method was widely applied and further developed in medical research (Gardner and Altman, 1986; Simon, 1986; Morgan, 1989). Confidence curves have been proposed for the presentation of study results more than 40 years ago. We wonder if there will be an increase in the frequency of applications of confidence curves in the medical literature, as was seen for the application of confidence intervals since the 1980s. One- and two-sided confidence curves are useful complements to the conventional methods of presenting the results of medical studies, especially in cases where different confidence levels are considered for medical decision making.

Acknowledgements We thank the Associate Editor for a number of helpful comments and for pointing out some errors in the formulas of earlier versions of the manuscript.

References

Altman, D. G., Machin, D., Bryant, T. M., and Gardner, M. J., eds. (2000). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. BMJ Books, London.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association* **56**, 246–249.

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* **28**, 783–798.

Bland, J. M. and Altman, D. G. (1994). One and two sided tests of significance. *British Medical Journal* **309**, 248.

Borenstein, M. (1994). The case for confidence intervals in controlled trials. *Controlled Clinical Trials* **15**, 411–428.

Cascinelli, N., Morabito, A., Santinami, M., MacKie, R. M., Belli, F., on behalf of the WHO Melanoma Programme (1998). Immediate or delayed dissection of regional nodes in patients with melanoma of the trunk: A randomised trial. *The Lancet* **351**, 793–796.

Chen, C. and Tipping, R. W. (2002). Confidence interval of a proportion with over-dispersion. *Biometrical Journal* **44**, 877–886.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics* **29**, 357–372.

Dunnett, C. W. and Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine* **15**, 1729–1738.

Folks, J. F. (1981). *Ideas of Statistics*. Wiley, New York.

Foster, D. A. and Sullivan, K. M. (1987). Computer program produces p-value graphics (Letter). *American Journal of Public Health* **77**, 880–881.

Gardner, M. J. and Altman, D. G. (1986). Confidence intervals rather than P values: Estimating rather than hypothesis testing. *British Medical Journal* **292**, 746–750.

Hung, H. M. J. and O'Neill, R. T. (2003). Utilities of the P-value distribution associated with effect size in clinical trial. *Biometrical Journal* **45**, 659–669.

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Köhne, K. (1997). The behavior of the P-value when the alternative hypothesis is true. *Biometrics* **53**, 11–22.

ICH E9 Expert Working Group (1999). ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials. *Statistics in Medicine* **18**, 1905–1942.

Kempthorne, O. and Folks, J. F. (1971). *Probability, Statistics, and Data Analysis*. Iowa State University Press, Ames.

Knottnerus, J. A. and Bouter, L. M. (2001). The ethics of sample size: Two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology* **54**, 109–110.

Lui, K.-J. (2000). Confidence intervals of the simple difference between the proportions of a primary infection and a secondary infection, given the primary infection. *Biometrical Journal* **42**, 59–69.

Lui, K.-J. and Lin, C.-D. (2003). A revisit on comparing the asymptotic interval estimators of odds ratio in a single 2×2 table. *Biometrical Journal* **45**, 226–237.

Mau, J. (1988). A statistical assessment of clinical equivalence. *Statistics in Medicine* **7**, 1267–1277.

Miettinen, O. S. (1985). *Theoretical Epidemiology.* Wiley, New York.

Morgan, P. P. (1989). Confidence intervals: From statistical significance to clinical significance. *Canadian Medical Association Journal* **141**, 881–883.

Moyé, L. A. (1998). P-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology* **8**, 351–357.

Moyé, L. A. (2000). Alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Statistics in Medicine* **19**, 767–779.

Poole, C. (1987a). Beyond the confidence interval. *American Journal of Public Health* **77**, 195–199.

Poole, C. (1987b). Confidence intervals exclude nothing. *American Journal of Public Health* **77**, 492–493.

Scherb, H. and Brüske-Hohlfeld, I. (1993). The exact confidence curve. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **24**, 152–163.

Schweder, T. and Spjotvoll, E. (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Shakespeare, T. P., Gebski, V. J., Veness, M. J., and Simes, J. (2001). Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *The Lancet* **357**, 1349–1353.

Sidney, S., Petitti, D. B., Quesenberry, C. P., Jr., Klatsky, A. L., Ziel, H. K. and Wolf, S. (1996). Myocardial infarction in users of low-dose oral contraceptives. *Obstetrics & Gynecology* **88**, 939–944.

Simon, R. (1986). Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine* **105**, 429–435.

Smith, A. H. and Bates, M. N. (1992). Confidence interval analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology* **3**, 449–451.

Sullivan, K. M. and Foster, D. A. (1990). Use of the confidence interval function. *Epidemiology* **1**, 39–42.