2001

Consensus evidence evaluation in resuscitation research: analysis of Type I and Type II errors

Charles F. Babbs
Purdue University, babbs@purdue.edu

Consensus evidence evaluation in resuscitation research: analysis of Type I and Type II errors

Charles F. Babbs, M.D., Ph.D.* [1]

* Department of Basic Medical Sciences, Purdue University, 1246 Lynn Hall, West Lafayette, IN 47907-1246, USA

1 Member of the ACLS Subcommittee and Chair of the Research Working Group, Emergency Cardiovascular Care Programs, American Heart Association.

ABSTRACT

Objective: This paper addresses the following statistical question: 'if genuine improvements in cardiopulmonary resuscitation (CPR) were discovered that doubled the probability of resuscitation success in a series of randomized clinical trials, would they be recognized and incorporated into consensus guidelines?'

Methods: Statistical powers for hypothetical individual clinical trials comparing experimental and control CPR were computed as a function of the study *N* when the true probabilities for immediate survival, 24 h survival, and discharge survival in the experimental group were twice those in the control group. Next, the binomial distributions describing the numbers of statistically significant studies in a series of equally powered trials of the same intervention were determined. These were compared with varying criteria for consensus among expert reviewers, expressed in terms of the number of 'positive' studies showing a statistically significant difference that reviewers would require before approving the experimental method.

Results: False-negative evaluations (i.e. failures to approve a technique that actually doubled survival) were extremely common under a wide range of realistic assumptions and consensus criteria, especially when simulated long-term survival data were considered. Similar methods showed that false-positive evaluations would be extremely rare, provided that at least two of the clinical trials in a series showed a statistically significant benefit of the experimental method.

Conclusions: Optimization of evidence evaluation can and should be carried out to make better use of available data in creating resuscitation guidelines. One simple approach is the 'two and one quarter test': if at least two well-conducted studies in a series are significantly positive ($P<0.05$) comprising at least one-quarter of all studies in the series, a positive effect can be inferred with small Type I and Type II errors. In addition, greater reliance on modern, unbiased methods such as cumulative meta-analysis is needed to increase the sensitivity of evidence evaluation for detecting useful innovations in resuscitation.

Keywords: Cardiopulmonary resuscitation; Guidelines; Human experimentation; Clinical trials; Meta-analysis; Standards

Resuscitation 51 (2001) 193–205

1. Introduction

Cardiopulmonary resuscitation (CPR) is unusual among medical treatments in that clinical practice is determined largely by national or international guidelines and to a much lesser extent by the judgment of individual clinicians. Such standardization minimizes chaos in highly emergent situations, but also discourages innovation. Improvements must be blessed by guideline writing committees, which tend to follow implicit consensus criteria. The guidelines that direct the efforts of thousands of individuals worldwide in lifesaving efforts are formulated by committees of volunteer experts in organizations such as the European Resuscitation Council and the American Heart Association [1]. These experts are typically medical professionals who rely on traditional methods of literature review, prior experience, and clinical judgment to arrive at a consensus, to which the fewest committee members can strenuously object. A potential improvement in resuscitation guidelines is proposed, and consensus is achieved after review, debate, and synthesis of evidence from various research studies [2]. Optimization of lifesaving efforts around the world is dependent on accurate outcomes of this process.

In the rare situations, in which there is a large number of consistently positive randomized clinical trials, consensus favoring change is easy. This situation, however, is highly unlikely in the domain of cardiopulmonary resuscitation. Compared with clinical trials of promising new drugs that are supported by large multinational corporations, trials of resuscitation techniques are underfunded and data poor. Even in a well-investigated field such as treatment of myocardial infarction [3,4], years, even decades, may pass as clinical trials accrue, yielding a mixture of seemingly 'conflicting' positive and neutral studies. In the domain of resuscitation, there has never been a method or technique supported by overwhelmingly positive data from multiple randomized clinical trials. Yet guideline writers must create guidelines anyway, striving to make the most efficient use of available data.

In this sense, one can think of the evidence evaluation process as a diagnostic test to detect the presence of potential improvements in CPR, for which the concepts of sensitivity and specificity come into play. Ideal guideline writers would avoid both false-positive evaluations (concluding a guideline change is beneficial when in fact it is not) and false-negative evaluations (concluding a guideline change is of no benefit, when in fact it would be). If the process is not specific, rescuers and their instructors will be burdened by needless, ineffective changes. If the process is not sensitive, life-saving improvements in resuscitation technique will be missed.

Lack of specificity corresponds to a 'Type I' statistical error or a false-positive conclusion. Lack of sensitivity corresponds to a 'Type II' statistical error or a false-negative conclusion [5]. In this paper, we shall examine the Type I and Type II errors inherent in consensus evidence evaluation in the field of resuscitation. Understanding of the underlying mathematical and statistical realities can lead to better criteria for evaluating potential improvements in resuscitation technique.

2. Materials and methods

2.1. *Approach*

This paper presents a family of thought experiments to simulate the evidence-based review of series of randomized clinical trials, in which there is a known true difference in resuscitation success. Of particular interest is the probability of reaching an incorrect consensus decision as a function of the numbers of patients in individual clinical trials and the criteria that are used to reach 'consensus'. In this way, one can examine the consequences of different evidence evaluation strategies. Nomenclature for the simulations is summarized in Table 1. All computations can be made readily on a routine Microsoft Excel spreadsheet, using arithmetic operators and functions for the square root, the binomial distribution, the normal distribution, and the inverse normal distribution.

Table 1 Nomenclature

Variables	Definitions
$B(s, t, p)$	Cumulative binomial distribution evaluated for s successes in t trials with probability p
$F^{-1}(x, \mu, \sigma)$	Inverse normal distribution for probability x, mean μ and standard deviation σ
$F(x, \mu, \sigma)$	Cumulative normal distribution evaluated at x, having mean μ and standard deviation σ
p	Observed proportion of successful outcomes
Δp	Difference in observed proportions of survivors (experimental − standard)
N	Total number of patients in a study
$P(n/m)$	Probability of obtaining n positive studies in a series of m replicated studies with the same underlying values of π_1 and π_2
Greek letters	
α	Type I error for a particular study
β	Type II error for a particular study
π	True probability of successful outcome
σ	Standard deviation of a data set
Subscripts	
1	Standard CPR
2	Experimental CPR

The general approach is first to determine the statistical power of hypothetical individual clinical trials in which the experimental resuscitation technique actually doubles survival. The statistical power of a trial is the probability of finding a statistically significant difference experimentally when a given true difference in survival between the experimental and control groups exists. Next, the binomial distribution for the number of significant positive studies in a series of similarly powered trials is determined. This distribution can be compared with the number of significant positive studies that would be required for approval by a review committee operating under various criteria for consensus.

A strict committee would require a large number of significant positive trials; a more lenient committee would require a smaller number of positive trials. When the number of studies exceeds the consensus criterion, a true-positive recommendation will result. When the number of studies falls short of the consensus criterion, a false-negative recommendation, or Type II error, will result. In this way, one can study the frequencies of Type II errors under a variety of plausible conditions typical in the field of resuscitation. These can be compared with the frequencies of Type I errors under similar consensus criteria when there is, in fact, no difference between experimental and control treatments. In this way, one can explore when errors in evidence evaluation are likely to occur and in turn suggest strategies to minimize them.

2.2. *Describing an individual study*

Consider a single randomized clinical trial in which a new resuscitation method is compared with an old one. The experimental method might involve a new form of thoraco-abdominal compression, a new method of ventilation, a new drug, a new defibrillation waveform, or a new sequence of live-saving maneuvers. Since the purpose of resuscitation is the restoration of life, clinical trials ultimately focus on survival data [6–8]. Generally, one or more of three classical outcome measures is tabulated in such studies: return of spontaneous circulation (ROSC), 24 h survival, or survival to hospital discharge. The experimental CPR technique is judged superior if there is a statistically significant difference in at least one of these dichotomous variables. Generally, by the time randomized clinical trials are organized and approved, results of animal studies and non-randomized human studies have ruled out substantial safety concerns regarding the new method. Hence, the key question becomes whether the new method produces greater or less overall resuscitation success than standard CPR.

The possible results of any such clinical trial can be described by the binomial sampling distributions shown in Fig. 1. These distributions show the range of possible outcomes of a particular clinical trial if the entire study were repeated a very large number of times in the same population. Here, the horizontal axis represents the difference between experimental and standard CPR in the proportion of immediate, 24 h, or discharge survivors. The vertical axis represents the probability density for the sampling distributions. ‘Probability density’ is scaled such that the area under each distribution is unity.

The left-hand distribution is computed for the null hypothesis that $\pi_1 = \pi_2$. That is, the true probability, π_2, of survival with experimental CPR is identical to the true probability, π_1, of

survival with standard CPR. The mean difference is zero, as expected, but there is substantial variation. The right-hand distribution is computed for the alternative hypothesis that $\pi_2 > \pi_1$. Here the mean difference is positive, reflecting the true-positive effect. There is, however, substantial variation in possible outcomes of the trial. In a small percentage of cases, the measured difference in survival could be negative, despite a true-positive effect.

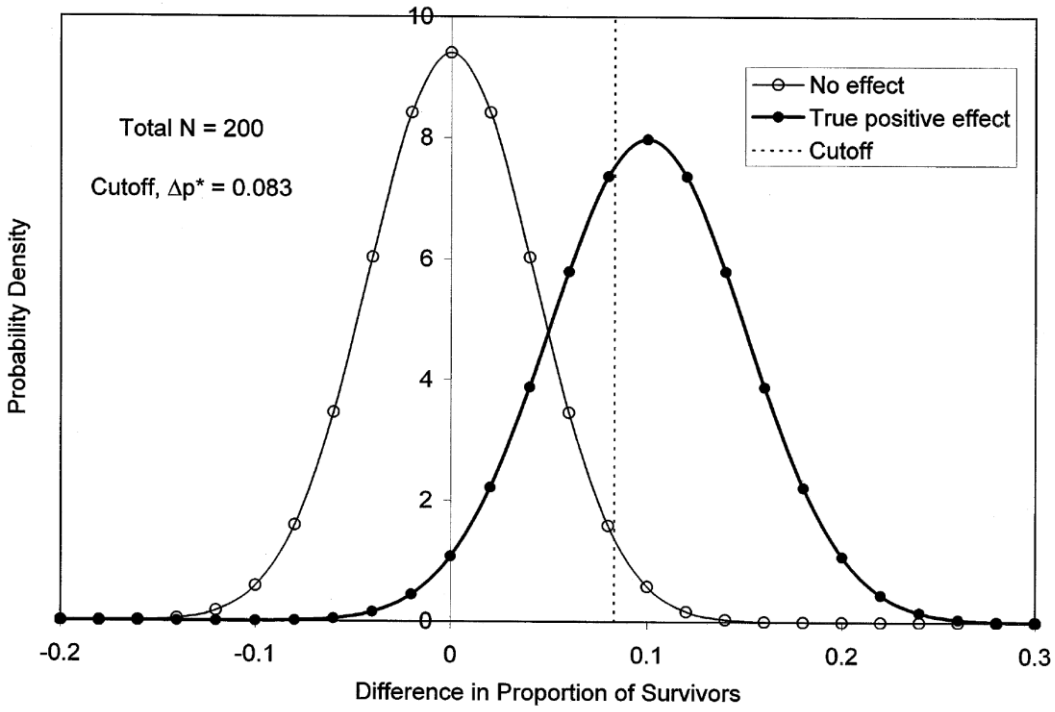


Fig. 1. Example of a distribution of measured differences in proportions of survivors under the null hypothesis ('No effect') and an alternative hypothesis ('True-positive effect') for a model of 24 h survival data in which standard CPR results in 10% survival and experimental CPR results in 20% survival. The modal difference is the expected 10% increase. Random variability in the results produces substantial overlap. Here, there are 100 patients in each group.

The curves in Fig. 1 represent a mathematical description of the possible results of a particular clinical trial, even though the actual difference in resuscitation success, $\pi_2 - \pi_1$, is constant. This inevitable variability of the binomial distribution occurs whenever one measures a dichotomous variable such as survival that describes each patient as a success or a failure, a plus or a minus, a 1 or a zero. As the number of patients in each group increases, the standard deviation of the difference in the proportion of survivors between experimental and control groups decreases according to the formula [9, 10]:

$$\sigma_{p_2-p_1} = \sqrt{\pi_1(1-\pi_1)/N_1 + \pi_2(1-\pi_2)/N_2} \qquad (1)$$

The dashed vertical line at the proportion, Δp^*, on the horizontal axis of Fig. 1 represents the critical value for statistical significance. Given the null hypothesis that there is no real difference between the groups ($\pi_2 = \pi_1$), the measured proportion of survivors in the experimental group would be greater than Δp^*only a small proportion of the time. The area under the curve representing 'No effect' (open circles) to the right of Δp^* is related to the Type I error, α, which is the probability of concluding a significant effect of treatment exists (either positive or negative), when in fact there is no difference. This area is equal to $\alpha/2$ (typically, 0.025) for two-sided or two-tailed tests of significance. According to the usual decision rules for statistical significance, a study is considered 'positive' if the observed success in the experimental group is greater than Δp^*. The study is considered neutral or 'negative' if the observed proportion of survivors in the experimental group is less than Δp^*.

The area under the curve representing a 'True-positive effect' (filled circles) to the left of Δp^* is the Type II error, denoted β, which is the probability of concluding there is no difference when in fact there is a difference. The area under the same curve to the right of Δp^*, or $1 - \beta$, is the power of the study. The power is the probability of obtaining a statistically positive result when a true treatment effect is present ($\pi_2 > \pi_1$).

According to the difference in proportion test [9, 11, 12], the critical difference in measured proportions, Δp^*, required for statistical significance can be computed from the inverse normal distribution function, F^{-1}, as

$$\Delta p^* = F^{-1}(1 - \alpha/2,\ 0,\ \sigma_{p_1 - p_2}) \tag{2}$$

Here, α is the acceptable Type I error for a two-tailed test of significance, and σ_{p1-p2} is the standard deviation of the difference in proportions, assuming the null hypothesis. For example, if $1 - \alpha/2$ is 0.975, then Δp^* is the 97.5th percentile of a normal distribution with a mean of zero and a standard deviation of σ_{p1-p2} . The value of σ_{p1-p2} is obtained exactly by evaluating the expression of Eq. (1) for $\pi_2 = \pi_1$. For the situation in which there is a true effect and $\pi_2 > \pi_1$, errors in statistical inference are Type II errors. For such a study, the Type II error, β, is given by the cumulative normal distribution function, F, as

$$\beta = F(\Delta p^*,\ \pi_2 - \pi_1,\ \sigma_{p_1 - p_2}) \tag{3}$$

This is the area from negative infinity to Δp^* under a normal distribution with mean $\pi_2 - \pi_1$ and standard deviation σ_{p1-p2} . In turn, the power of the study, or the probability that the study will yield a significant positive result, is $1 - \beta$.

2.3. *Describing likely treatment effects in resuscitation trials*

Treatment effects for typical resuscitation studies are easily imagined. For standard CPR, the approximate frequency of successful ROSC is about 25%, that of 24-h survival is about 10%, and that of hospital discharge is about 5% [13–18]. These values represent typical control outcome measures. Hypothetical large and small improvements in survival as the result of an experimental technique are indicated in Table 2. A large treatment effect is modeled as one that doubles positive outcome measures. For the purpose of comparison, a small treatment effect is modeled as one that improves outcome measures by 20%.

Table 2 Hypothetical positive treatment effects for simulation of consensus decision making, expressed in terms of true probabilities π, of successful resuscitation by various measures

	Discharge survival	24 h survival	ROSC[a]
Standard CPR control	$\pi_1 = 0.05$	$\pi_1 = 0.10$	$\pi_1 = 0.25$
Large improvement in CPR	$\pi_2 = 0.10$	$\pi_2 = 0.20$	$\pi_2 = 0.50$
Small improvement in CPR	$\pi_2 = 0.06$	$\pi_2 = 0.12$	$\pi_2 = 0.30$

[a] Return of spontaneous circulation.

2.4. *Describing series of clinical trials*

Imagine an ideal world in which a particular resuscitation study could be replicated many times independently in the same general population of patients. Suppose each of the replicated studies consists of an experimental and a control group, for which survival data after resuscitation are reported. For present purposes, it will suffice to let all replications have the same total number of patients, N, and for simplicity to let the N patients in each study be equally divided between experimental and standard CPR. It then becomes straightforward to explore N as a parameter.

Suppose there is a true treatment effect such that $\pi_2 > \pi_1$, as in Fig. 1 (filled circles). Table 3 presents the formulae for calculating the probabilities of obtaining a given number of statistically significant, 'positive' studies in such a series of replications. These probabilities are derived from the basic independence, product, and addition rules of probability theory [19], and from the definition of the power of a study $(1 - \beta)$. The left hand column in Table 3 presents the total number of studies in a series. Each column to the right indicates a given number of statistically

significant studies in a series, ranging from none of the studies in the series being significant to all of the studies in the series being significant.

Table entries are the probabilities of obtaining the indicated number of statistically significant trials in a particular column under the assumed conditions. For example, if there is only one trial in the series, the probability that it will not be significant, despite a true effect is the Type II error, β, and the probability that it will be significant is the power, 1 − β. Accordingly, for a series containing only one study, the probability of obtaining zero positive studies in the series is β, and the probability of obtaining one positive study in the series is 1 − β. If there are two trials, the probability that both will be negative is β^2 and the probability that both will be positive is $(1-\beta)^2$. The probability that one will be positive and one will be negative is β(1 − β), but there are two ways this can happen. Hence, the probability that one of the two studies will be statistically positive is 2β(1 − β). This process can be continued for larger series, leading to the remaining entries in Table 3. The binomial nature of these probabilities is well known [20, 21].

Table 3 Probabilities of various outcomes for series of independently replicated studies with the same statistical power (1 − β)

Total number of studies in the series	Number of statistically significant studies in the series						
	0	1	2	3	4	5	6
1	β	$(1-\beta)$					
2	β^2	$2\beta(1-\beta)$	$(1-\beta)^2$				
3	β^3	$3\beta^2(1-\beta)$	$3\beta\ (1-\beta)^2$	$(1-\beta)^3$			
4	β^4	$4\beta^3(1-\beta)$	$6\beta^2(1-\beta)^2$	$4\beta(1-\beta)^3$	$(1-\beta)^4$		
5	β^5	$5\beta^4(1-\beta)$	$10\beta^3(1-\beta)^2$	$10\beta^2(1-\beta)^3$	$5\beta\ (1-\beta)^4$	$(1-\beta)^5$	
6	β^6	$6\beta^5(1-\beta)$	$15\beta^4(1-\beta)^2$	$20\beta^3(1-\beta)^3$	$15\beta^2(1-\beta)^4$	$6\beta\ (1-\beta)^5$	$(1-\beta)^6$

2.5. *Describing consensus among evidence evaluators*

Although consensus is actually achieved by human judgment and group dynamics, members of the group seem to follow unconscious mathematical rules that can be used to create an operational definition of consensus for the purpose of the present analysis. This process has been described in the statistical literature as ‘vote counting’ by Hedges, Ingram, and other workers [10, 20–22]. In the vote-counting paradigm, each study ‘casts a vote’ in favor of the experimental intervention if it shows significant positive results. The study casts a vote against the intervention otherwise. The reviewers conclude there is a genuine treatment effect when there is a certain proportion or more of positive votes.

Vote counting is probably the most common decision procedure used in traditional research reviewing. If the proportion of positive ‘votes’ is large, then the treatment under investigation presumably has an effect. For example, a group of evaluators might consider evidence compelling if 75% or more of trials are significantly positive. The group might then reach a consensus that the innovation under study is effective and worthy of recommendation.

Review committees are at liberty to select any such level, which may vary with the total number of studies. At the 75% level, it seems to be a safe bet to agree with a positive conclusion. However, when only one-half of the studies show a statistically significant positive effect, and one-half do not, many authorities tend to conclude, often erroneously, that the research is inconclusive and 'more research is needed' [10].

For the purposes of the present analysis, we shall define a consensus threshold, c, as the minimum proportion of statistically positive studies in a series that is judged sufficient to justify a strong recommendation of a new procedure. For example, a 3/4 consensus threshold would describe the thinking of a group convinced by four of four or three of four positive studies, but not by two of four. For larger series of replicated studies, the consensus threshold is just the overall fraction or percentage of statistically significant positive studies.

2.6. *False-positive consensus evaluations*

In situations in which there is no real treatment effect, evaluators can come to either a true-negative or a false-positive consensus. The probability of reaching an incorrect, false-positive consensus in a series of replicated studies having the same two-tailed Type I error, α, is easily computed. If there is only one study in the series, the probability of a false-positive result is $\alpha/2$. If there are two studies, A and B, and if a 2/2 consensus is required, the probability that both A and B are falsely positive, which is equal to the probability of a false-positive consensus, is $(\alpha/2)(\alpha/2) = \alpha^2/4$. For $\alpha = 0.05$, this value is 0.000625. If only a 1/2 consensus is required, the probability is $(\alpha^2/4) + 2(\alpha/2)[1-(\alpha/2)]$. This expression corresponds to the combined probabilities for A+/B+, or B+/A−, or B−/A+, where the superscript + means that the results of study were positive for two-sided significance level α, and the superscript − means that they were not. For $\alpha = 0.05$, the probability for a 1/2 consensus or better is 0.049375. In general, the probability of reaching an incorrect false-positive n/m consensus is

$$P_{\mathrm{fp}}(n/m) = \left(\frac{\alpha}{2}\right)^m + \binom{m}{1}\left(\frac{\alpha}{2}\right)^{m-1}\left(1-\frac{\alpha}{2}\right) + \binom{m}{2}\left(\frac{\alpha}{2}\right)^{m-2}\left(1-\frac{\alpha}{2}\right)^2 + \ldots + \binom{m}{n}\left(\frac{\alpha}{2}\right)^{m-n}\left(1-\frac{\alpha}{2}\right)^n \qquad (4)$$

where

$$\binom{m}{k}$$

refers to the number of combinations of m things taken k at a time [23].

2.7. *False-negative consensus evaluations*

In situations in which there is a true treatment effect, evaluators can come to either a true-positive or a false-negative consensus. In these situations, the probability of obtaining a false-negative consensus is 1 minus the probability of obtaining a true-positive consensus. The probability of obtaining a true-positive consensus for a given consensus threshold may be computed from the terms in Table 3, working from right to left by rows, for example, as follows:

$$P_{\text{tp}}(4/4) = (1-\beta)^4$$

$$P_{\text{tp}}(5/6) = (1-\beta)^6 + 6\beta(1-\beta)^5$$

and, in general,

$$P_{\text{tp}}(n/m) = (1-\beta)^m + \binom{m}{1}\beta(1-\beta)^{m-1} + \binom{m}{2}\beta^2(1-\beta)^{m-2} + \ldots + \binom{m}{m-n}\beta^{m-n}(1-\beta)^n. \quad (5a)$$

Eq. (5a) can be expressed more compactly in terms of the cumulative binomial distribution function $B(s, t, p)$ for s successes in t trials with probability p. This function is available in Microsoft Excel and was used for spreadsheet computations. Using this function, Eq. (5a) is equivalent to

$$P_{\text{tp}}(n/m) = B(m-n,\ m,\ \beta) \quad (5b)$$

In turn, the probabilities of coming to a false-negative consensus for each consensus threshold may be computed as

$$P_{\text{fn}}(n/m) = 1 - P_{\text{tp}}(n/m) = 1 - B(m-n,\ m,\ \beta) \quad (6)$$

2.8. *Design of simulations*

Using the forgoing concepts, one can perform a series of thought experiments to explore the outcome of the evaluation of a series of studies that compare an experimental treatment with a standard treatment as a function of three key variables. These include the number of patients in each study, the true difference, if any, in resuscitation success, and the consensus threshold. Of special interest are the probabilities of false-positive and false-negative evaluations as functions of these variables.

3. Results

3.1. *False-positive consensus decisions*

Table 4 shows values computed from Eq. (4) for series of four, six, eight, and ten studies, assuming $\alpha = 0.5$. Each column represents the total number of studies and each row represents the minimum number of positive studies in the series for a positive consensus. Table entries are probabilities that a positive consensus would be reached when there is in fact no difference between treatment groups, i.e. $\pi_2 = \pi_1$. In this situation, all errors of research evaluation and synthesis are Type I errors. Whenever there are more than two positive studies in a series, the values in Table 4 are quite small. For series of four or six studies, the presence of only two positive studies excludes the null hypothesis at the P = 0.01 level (bold font). That is, the probability that the null hypothesis is correct is less than 1%. For series of eight or ten studies the presence of only three positive studies excludes the null hypothesis at the P = 0.01 level. Thus, as previously described [10, 20], only two or three significant positive studies are required to reject the null hypothesis.

Table 4 Probabilities of false-positive consensus decisions based on a series of replicated studies in which there is no true difference and the two-tailed Type 1 error, α, of each study is 0.05

Minimum number of positive studies needed for consensus	Total number of studies in the series			
	4	6	8	10
1	0.096 312 109 4	0.140 931 699 0	0.183 348 196 3	0.800 941 122 9
2	**0.003 626 171 9**	**0.008 767 345 0**	0.015 829 877 6	0.024 611 502 0
3	0.000 061 328 1	0.000 295 271 0	**0.000 796 182 4**	**0.001 643 170 0**
4	0.000 000 390 6	0.000 005 627 4	0.000 025 223 6	0.000 072 685 8
5		0.000 000 057 4	0.000 000 513 4	0.000 002 215 3
6		0.000 000 000 2	0.000 000 006 5	0.000 000 047 0

These results are independent of the particular kinds of data or significance tests used in the studies. They are also independent of the number of patients, *N*, in any of the studies, because the *N* values of the individual studies are already taken into account in selecting the critical values for significance testing.

3.2. *False-negative consensus decisions*

For cases in which π_2 is greater than π_1, the errors in evidence evaluation are Type II errors. Here, the probabilities of false-negative consensus decisions are dependent on study *N* values. For simplicity in the present analysis, we assume the same *N* for all studies in the series. Fig. 2 illustrates probabilities for false-negative conclusions as a function of *N* in a variety of scenarios in which the true difference in survival for experimental CPR is twice that for standard CPR, i.e. $\pi_2 = 2\pi_1$. Three different control levels of survival are modeled as indicated in Table 2, corresponding to discharge survival, 24-h survival, and ROSC for typical standard CPR.

[Please continue on next page.]

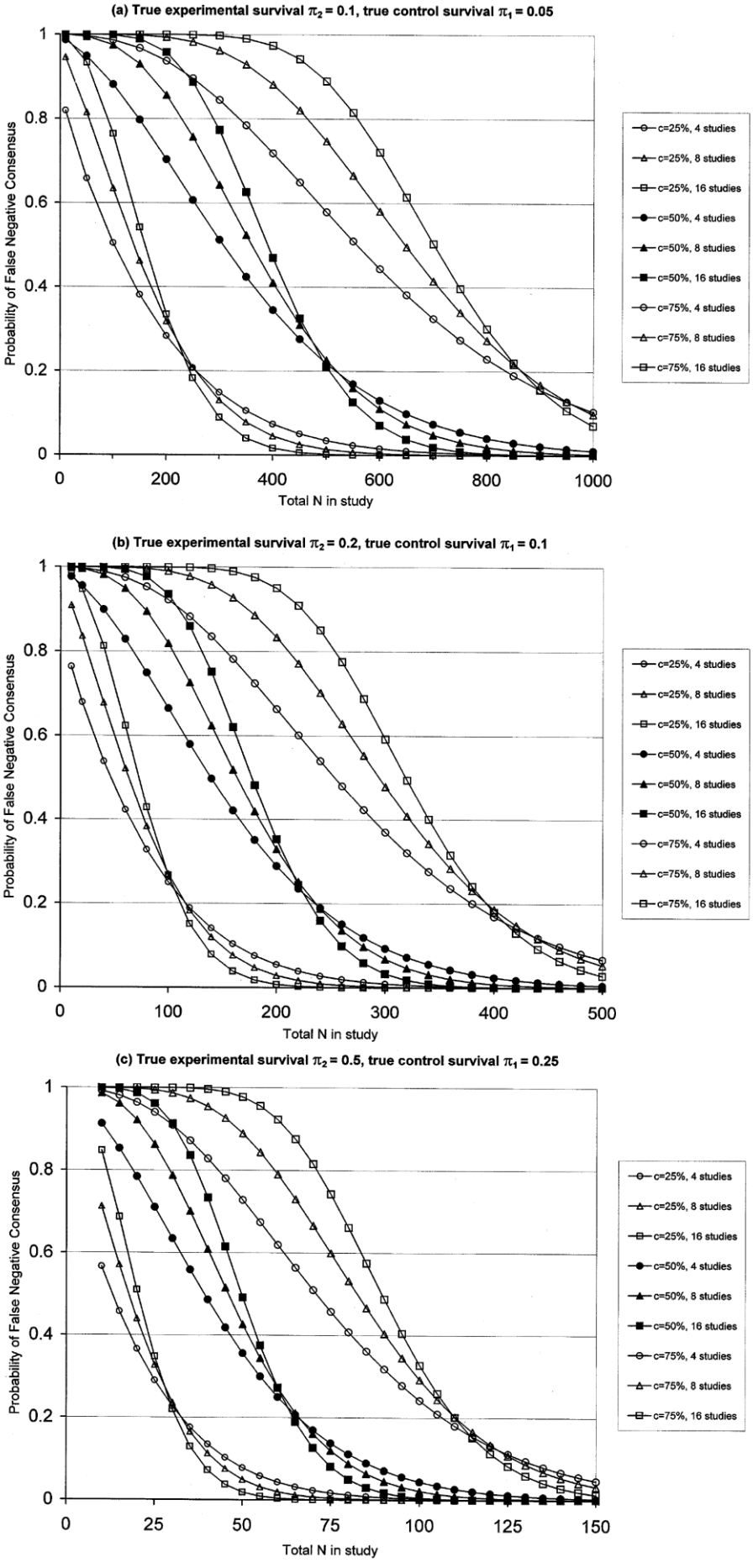


Fig. 2. Probabilities of reaching a false-negative consensus from series of replicated clinical trials, given an actual twofold improvement in resuscitation success. Separate charts describe scenarios simulating three different survival endpoints commonly measured in resuscitation research. Clustered groups of three curves represent possible consensus criteria. Solid symbols indicate that at least one-half of studies must show a significantly positive result with $\alpha = 0.05$. Smaller open symbols indicate that at least one-quarter of studies must show a significantly positive result. Larger open symbols indicate that at least three-quarters of studies must show a significantly positive result. Symbol types represent the number of studies to be evaluated: circles, four studies; triangles, eight studies; squares, 16 studies. (a) Stimulates discharge survival, (b) 24-h survival, (c) return of spontaneous circulation.

For each control survival level (a)–(c), the probabilities of false-negative evaluations are plotted as a function of the total number of patients, *N*, in each of the replicated studies. Keep in mind that for every case the true probability of survival for experimental CPR, π_2, is twice that of control CPR, π_1. The nine curves in each chart appear in groups of three. Each group represents a different consensus threshold. The middle group (solid symbols) represents a consensus threshold of 50%. That is, evaluators require 50% or more of the reviewed studies to show statistically significant results before reaching a consensus that the innovation under study is truly effective. The left-hand group of three curves (small open symbols) represents a less stringent consensus threshold of 25%. The right hand group of three curves (larger open symbols) represents a more conservative consensus threshold of 75%. The three curves within each group indicate the evaluation of four, eight, or 16 similar studies. Circles indicate a series of four studies, triangles a series of eight studies, and squares a series of 16 studies.

Fig. 2(a) shows results typical of long-term, discharge survival as an end-point. Long-term survival is the most valued end point in resuscitation research, based on the laudable desire of both patients and clinicians to eschew methods that restore circulation but prolong life only a few hours or days. Such methods would increase cost and suffering without increasing quality of life. Compared with the probabilities for false-positive evaluations in Table 4, the plotted probabilities for false negative evaluations in Fig. 2(a) are large. For small to medium sized studies, including fewer than 1000 patients, the chances of false-negative evaluations can be 50% or greater for long-term survival endpoints.

For reference, the mean study *N* values in current trials of experimental CPR is in the range of 300 total patients [16, 17]. Brown et al.'s review of negative studies in emergency medicine [12] found a mean *N* of 82 and range of 12–394 for the number of patients in both experimental and control groups. In this range of *N*, the probability of making a false-negative consensus evaluation of an experimental method that in fact doubles long-term discharge survival is very high.

The sensitivity of the evidence evaluation process is 1 minus the probability of reaching a false-negative consensus. For studies with *N* in the range of 50–500 patients, the sensitivity of evidence evaluation based on long-term survival data is roughly 50%. That is, the overall assessment of research literature is correct about half the time. ***For realistically conservative consensus criteria, requiring at least 50% of eight or more studies to be statistically significant, the ability of the evidence evaluation process to recognize an innovation that actually doubles long term survival is less than 10%.*** In order for the probability of false-negative consensus to become less than 10%, the studies must include 1000 or more patients each. The consensus threshold does have a large influence on the sensitivity of the evidence evaluation process, especially for small to medium sized studies. Only when total study *N* is greater than 1000 is the process insensitive to the consensus threshold.

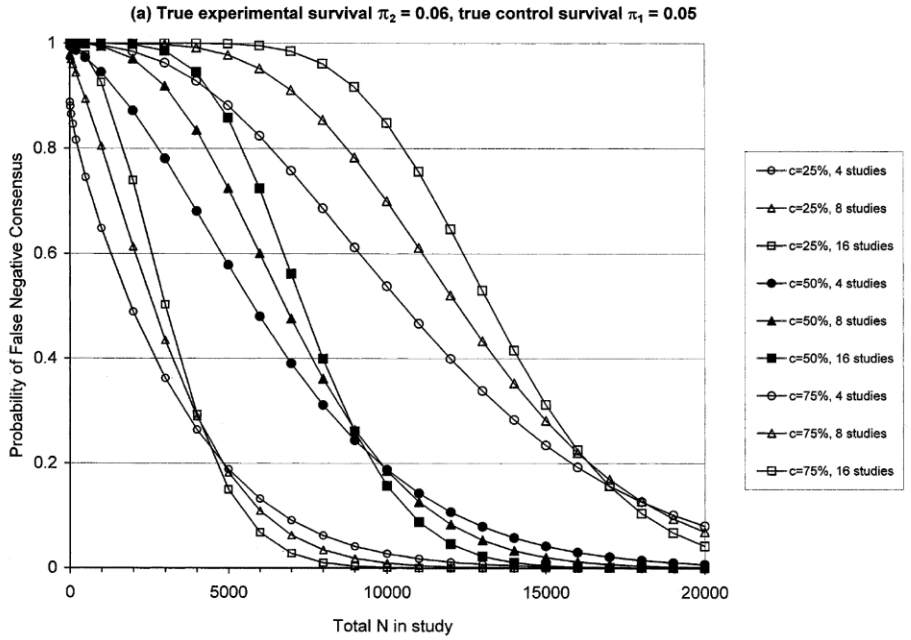
Results within each group of curves are also revealing and important. Consider the solid symbols in Fig. 2(a). These represent a consensus threshold of 50%. Note that when the curve height is greater than about 0.2, the probability of false-negative evaluation actually increases as the

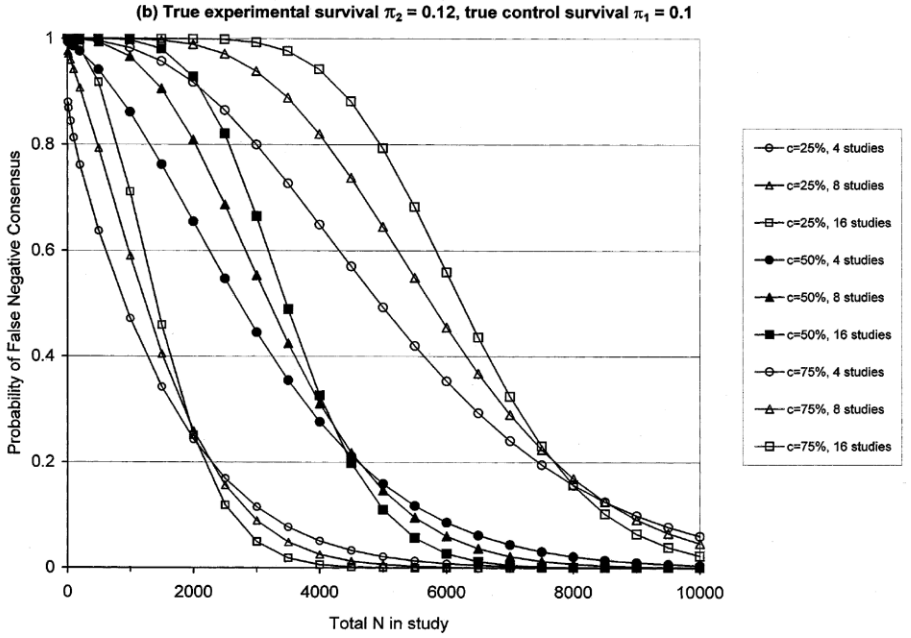
number of studies to be integrated increases from four to 16. Hedges and Olkin described this counterintuitive effect in 1980 [20]. We shall refer to it as the Hedges–Olkin paradox. The paradox is that, under certain conditions, the larger the number of studies conducted, the greater is the certainty of reaching the false conclusion that there is no effect. This paradox is a genuine consequence of the laws of probability.

The necessary conditions for the Hedges–Olkin paradox are common in resuscitation research and occur when the statistical power of the studies is less than the consensus threshold. The Hedges–Olkin paradox is true because, if a given study were replicated a large number of times, the proportion of replications with statistically significant results would be equal to the power. Consequently, if the power is less than the consensus threshold then, as more studies are carried out, the proportion of positive studies will tend to become less than the consensus threshold. This conclusion is also true if the study powers vary, but the average power is less than the consensus threshold ([21], p. 52). ***Thus, under conditions common in the field of resuscitation, the sensitivity of the consensus process to detect true-positive effects decreases as the number of studies reviewed increases!*** In turn, the probability that a research review reaches the correct decision tends toward zero as more research is carried out [20]!

This disturbing situation is improved slightly by choosing end points with greater π_1, such as 24-h survival or immediate survival after attempted resuscitation (ROSC). For typical 24-h survival with π_1 in the range of 0.1, the evaluation of studies with N values in the range 100–200 has a 50% sensitivity with a 50% consensus threshold (Fig. 2(b)). The Hedges–Olkin paradox is still in effect, however. When π_1 is in the range of 0.25, corresponding to the ROSC end-point, then consensus evidence evaluation is capable of detecting an effect with a sensitivity greater than 90% for typically powered resuscitation studies (Fig. 2(c)).

The results in Fig. 2 were computed for studies reflecting a large, two-fold, improvement in resuscitation success. Fig. 3 illustrates similar results for more modest, incremental improvements in survival of 20%. Similar patterns of findings occur as before. However, the N values required to achieve a given level of sensitivity are much larger. False-negative evaluations are much more likely with small N studies typical of resuscitation research. For the endpoint of long-term survival, the Hedges–Olkin paradox reigns supreme.





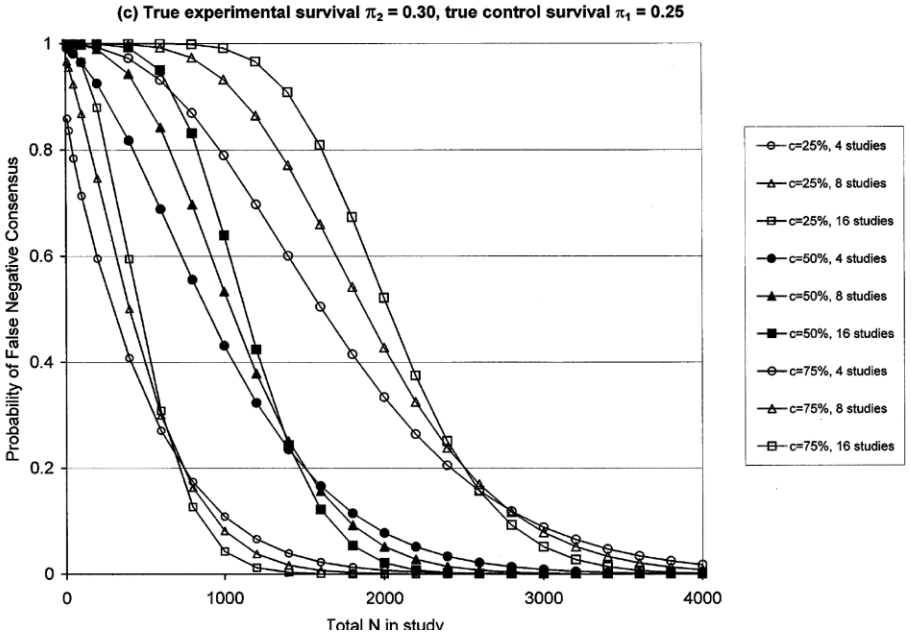


Fig. 3. Probabilities of reaching a false-negative consensus from series of replicated clinical trials given an actual 20% improvement in resuscitation success. Other details similar to Fig. 2. Note expanded scales for N.

Extremely large *N* values are required to obtain adequate sensitivity of the evaluation process — probably so large that cost would be prohibitive. Even using ROSC as the only endpoint, greater than 1000 patients in each study are needed. Only one recent resuscitation study has included this many patients [24]. Smaller positive effects are almost impossible to detect. These findings suggest that, in the field of resuscitation, traditional methods of research review and consensus development cannot lead to gradual improvement of guidelines by a series of small incremental steps. Only improvements with a large (e.g. two-fold or greater) impact upon outcome are likely to be adopted, and these only with good luck.

4. Discussion

The laws of probability and statistics have a lot to say about the way resuscitation research is conducted and the way resuscitation guidelines are created. Although studies of blood flow and hemodynamics deserve close scrutiny, the ultimate dependent variable in resuscitation is a dichotomous one—survival—putting investigators at the mercy of the binomial distribution.

The extreme pathophysiology of sudden cardiac death, often including long down times and severe underlying disease, limits both control survival rates (π_1) and possible improvements in survival rate ($\pi_2 - \pi_1$) even with the best of treatments. Small values of these parameters drive up the study *N* values needed to avoid both Type I and Type II errors in evidence evaluation.

Unfortunately, resuscitation has been an 'orphan research domain', underfunded by federal governments and by multinational drug companies. Hence, *N* values of most resuscitation studies are small. Under these conditions, the traditional concepts of research reviewing often cannot function reliably to recognize clinically significant improvements when they are discovered. The metaphor of 'weighing' the evidence that underlies traditional research synthesis seems intuitively fair. One simply stacks up the significantly positive studies in one hand and the non-significant, neutral studies in the other hand with some attention to their technical merits. If one stack is larger than the other, the decision seems obvious. This approach has deep roots in our system of justice and our normal approach to making choices in life by weighing 'pros' and 'cons'.

This vote counting approach, however, is mathematically erroneous. The error is revealed in Fig. 1, and has to do with the properties of dichotomous variables with binomial distributions. In the presence of a true effect, some studies will show no apparent difference, especially underpowered ones. Studies showing no statistically significant difference are not necessarily 'negatives' or 'cons'. One such study does not necessarily 'cancel' a significant positive study. It should be seen as part of a larger distribution.

A related misunderstanding is that conventional significance testing ensures accurate decisions about treatment effects. This reasonable sounding idea is only partly correct. Conventional significance testing only minimizes Type I errors. It does not in any way protect against Type II errors [12]. Indeed, over-reliance on conventional significance tests by requiring smaller and

smaller *P* values actually increases the number of Type II errors and reduces the overall accuracy of evidence evaluation.

In this setting, reliance on conventional methods of research reviewing and consensus will sacrifice many, perhaps most, genuine innovations in resuscitation technique. In particular, the current approach to guideline development would seem to obviate a series of successive incremental improvements (5% better ROSC here, 10% better ROSC there), which might have a substantial combined impact on overall success. Large, roughly two-fold improvements in resuscitation success, such as those reported by Sack et al. for interposed abdominal compression CPR [25] have been quite rare. As a result, major improvements in basic CPR have not appeared since its introduction in 1960 [26, 27]. In turn, researchers tend to become discouraged because their work, even when successful, never makes it into the guidelines.

Inspection of the results in Table 4 describing Type I errors, together with those in Figs. 2 and 3 describing Type II errors, leads to a simple approximate rule for eliminating many, but not all, systematic errors in evidence evaluation. This rule of thumb might be called the 'two and one quarter test'. If the same intervention is tested repeatedly in the same population, and if there are at least two well-conducted significant positive trials representing one-quarter or more of all trials performed, then one can conclude with reasonable accuracy that the innovation under study has a true-positive effect. This simple test can serve as a poor man's meta-analysis. One only has to count to two and to divide by four. This way of counting studies is much less stringent than the consensus criteria used by many research reviewers. Yet it is based soundly on the mathematical realities of the binomial distribution.

The two and one quarter test is of course an approximation. It remains a form of vote counting. All forms of vote counting include a systematic bias toward Type II errors [22], especially when studies are substantially underpowered or the experimental effects are small (Figs. 2 and 3). A better way to obtain both increased sensitivity and increased specificity of evidence evaluation is to conduct a formal meta-analysis, which does not suffer from the systematic bias of vote counting or from the Hedges–Olkin paradox. The details of meta-analysis are beyond the scope of the present paper. The reader is referred to the cited references for an introduction to this topic. The two and one quarter test suggested here, however, can be used to identify candidate interventions for formal meta-analysis.

Specific examples in need of meta-analysis in the field of resuscitation include interposed abdominal compression CPR (IAC-CPR) and active compression decompression CPR (ACD-CPR). Both methods have been shown in more than two randomized clinical trials to produce clinical benefit in human beings compared with standard CPR. In both cases, the positive trials constitute more than one-quarter of the randomized clinical trials performed [1]. Of course, analysis of these real world innovations is more complex than the simple thought experiments presented in the present paper. In particular, the same intervention was not necessarily tested in all studies, because the techniques for performing IAC-CPR and ACD-CPR have evolved and improved with time and experience.

Also, the studies of these techniques were not always replicated in the same populations. There are pre-hospital versus in-hospital trials. There are trials in different countries with different

health care systems by rescuers with different degrees of training [1]. Still, some of the variability in the results of multiple trials must have been caused by random variation of the binomial distribution, i.e. by luck. Typically, research reviewers tend to underestimate the 'normal' amount of random variation in binomial data and over-interpret apparent differences. A humorous account of this tendency is provided by Hunter and Schmidt [10]. For interventions such as IAC-CPR and ACD-CPR that pass the two and one quarter test, more sophisticated techniques than those of traditional consensus development are needed to improve the accuracy of the evidence evaluation process.

The techniques of meta-analysis, a general term for quantitatively combining evidence from related but independent studies, have recently become popular in the clinical literature. These methods have been well reviewed [10, 21 ,22, 28–31] and deserve serious consideration by guidelines writers. One important innovation is the technique of ***cumulative*** meta-analysis of outcome data, developed by Lau, Mosteller, Chalmers, and coworkers [3, 4]. Cumulative meta-analysis is defined as the performance of an updated meta-analysis every time a new trial appears. This approach simplifies the process of integrating data from multiple clinical trials, and makes it possible to pinpoint the earliest time when the combined results of clinical trials first achieve statistical significance. Importantly, meta-analysis eliminates the Hedges–Olkin paradox. Meta-analysis also tends to focus more attention on actual data, balancing the contributions of non-native English speakers, less domineering personalities, and less persistent advocates with their opposites on the evidence evaluation team.

5. Conclusions

The results of the present study suggest that evidence for new resuscitation guidelines need not be 'compelling' in the sense that a majority or a super-majority of published studies are statistically significant. Many positive life-saving innovations would probably never be implemented under these conditions. Indeed, they would only be adopted if the proportion of positive studies were much greater than normally expected on the basis of probability theory. This means that the measured treatment effects ($p_2 - p_1$) would have to be substantially greater than the true treatment effects ($\pi_2 - \pi_1$) — a matter of luck rather than merit.

A long term consequence of this situation is that subsequent clinical results of a few lucky techniques would necessarily tend to be 'disappointing', i.e. closer to the true average outcome than to the initial lucky outcome. In turn, observers would tend to grow cynical about future innovations, perhaps requiring even more conservative consensus criteria and creating a self-fulfilling prophecy. Unknowingly at the mercy of the Hedges–Olkin paradox, guideline writers would increasingly view resuscitation research as muddled, unproductive, and conflicting. Change would be put off once more, with yet another call for more data. A better approach would be a more realistic one, permitting modest incremental improvements based on the two and one quarter test, followed by unbiased meta-analysis of available results.

Ideally, consensus guidelines should be a conduit for the prompt transfer of effective innovations from research centers to front line patient care.

References

[1] Cummins RO. Advanced Cardiac Life Support. Emergency Cardiovascular Care Programs. Dallas, TX: American Heart Association, 1997.

[2] Cummins RO. Introduction to the international guidelines 2000 for CPR and ECC. Circulation 2000;102(suppl. I):I-1–I-11.

[3] Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 1992;327:248–54.

[4] Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. J Am Med Assoc 1992;268:240–8.

[5] Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 'negative' trials. N Engl J Med 1978;299:690–4.

[6] Cummins RO, Chamberlain D, Hazinski MF, et al. Recommended guidelines for reviewing, reporting, and conducting research on in-hospital resuscitation: the in-hospital 'Utstein style'. Circulation 1997;95:2213–39.

[7] Idris AH, Becker LB, Ornato JP, et al. Utstein-style guidelines for uniform reporting of laboratory CPR research. A statement for healthcare professionals from a task force of the American Heart Association, the American College of Emergency Physicians, the American College of Cardiology, the European Resuscitation Council, the Heart and Stroke Foundation of Canada, the Institute of Critical Care Medicine, the Safar Center for Resuscitation Research, and the Society for Academic Emergency Medicine. Writing Group. Circulation 1996;94:2324–36.

[8] Chamberlain D, Cummins RO. Recommended guidelines for uniform reporting of data from out-of-hospital cardiac arrest: the 'Utstein style'. European Resuscitation Council, American Heart Association, Heart and Stroke Foundation of Canada and Australian Resuscitation Council. Eur J Anaesthesiol 1992;9:245–56.

[9] Dunn OJ. Basic Statistics: A Primer for the Biomedical Sciences. New York: Wiley, 1964:113.

[10] Hunter JE, Schmidt FL. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. Newbury Park, CA: Sage Publications, 1990:592.

[11] O'Brien PC, Shampo MA. Statistics for clinicians 8 — comparing two proportions: the relative deviate test and chi-square equivalent. Mayo Clin Proc 1981;56:513–5.

[12] Brown CG, Kelen GD, Ashton JJ, Werman HA. The beta error and sample size determination in clinical trials in emergency medicine. Ann Emerg Med 1987;16:183–7.

[13] Eisenberg MS, Hallstrom A. Long term survival after out-of-hospital cardiac arrest. N Engl J Med 1982;306:1340–3.

[14] Kerber RE, Ornato JP, Brown DD, et al. Guidelines for cardiopulmonary resuscitation and emergency cardiac care. J Am Med Assoc 1992;268:2171–302.

[15] Nichol G, Stiell IG, Laupacis A, Pham B, De Maio VJ,Wells GA. A cumulative meta-analysis of the effectiveness of defibrillator-capable emergency medical services for victims of out-of-hospital cardiac arrest. Ann Emerg Med 1999;34:517–25.

[16] Sack JB, Kesselbrenner MB. Hemodynamics, survival benefits, and complications of interposed abdominal compression during cardiopulmonary resuscitation. Acad Emerg Med 1994;1:490–7.

[17] Tucker KJ, Savitt MA, Idris A, Redberg RF. Cardiopulmonary resuscitation. Historical perspectives, physiology, and future directions. Arch Intern Med 1994;154:2141–50.

[18] Woodhouse SP, Cox S, Boyd P, Case C, Weber M. High dose and standard dose adrenaline do not alter survival, compared with placebo, in cardiac arrest. Resuscitation 1995;30:243–9.

[19] Jekel JF, Elmore JG, Katz DL. Epidemiology Biostatistics and Preventive Medicine. Philadelphia:W.B.SaundersCompany,1996. 297 pp.

[20] Hedges LV, Olkin I. Vote-counting methods in research synthesis. Psychol Bull 1980;88:359–69.

[21] Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press, 1985.

[22] Wolfe FM. Meta-analysis: quantitative methods for research synthesis. In: Lewis-Beck MS, editor. Quantitative Applications in the Social Sciences, Vol. 59. Newbury Park, CA: Sage Publications, 1986. 65 pp.

[23] Feller W. An Introduction to Probability Theory and Its Applications, Vol. 1. New York: Wiley, 1957. 461 pp.

[24] Plaisance P, Adnet F, Vicaut E, et al. Benefit of active compression-decompression cardiopulmonary resuscitation as a prehospital advanced cardiac life support. A randomized multicenter study. Circulation 1997;95:955–61.

[25] Sack JB, Kesselbrenner MB, Bregman D. Survival from in-hospital cardiac arrest with interposed abdominal counterpulsation during cardiopulmonary resuscitation. J Am Med Assoc 1992;267:379–85.

[26] Kouwenhoven WB, Jude JR, Knickerbocker GG. Closed-chest cardiac massage. J Am Med Assoc 1960;173:1064–7.

[27] Weil MH, Tang W. Cardiopulmonary resuscitation: a promise as yet largely unfulfilled. Disease-a-month 1997;43:429–501.

[28] Mulrow CD. Rationale for systematic reviews. Br Med J 1994;309:597–9.

[29] Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. Stat Med 1987;6:315–28.

[30] Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987;316:450–5.

[31] Thacker SB. Meta-analysis. A quantitative approach to research integration. J Am Med Assoc 1988;259:1685–9.