4-2007

Statistical analysis of joint short-term and long-term survival in resuscitation research

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

Statistical analysis of joint short-term and long-term survival in resuscitation research

Charles F. Babbs[a,b]

[a] Department of Basic Medical Sciences, Purdue University; 1246 Lynn Hall, West Lafayette, IN 47907-1246, USA

[b] Indiana University School of Medicine, Indianapolis, IN, USA

Address for correspondence and proofs:
Charles F. Babbs, MD, PhD
1426 Lynn Hall
Purdue University
West Lafayette IN 47907-1246, USA
E-mail: babbs@purdue.edu
Voice: 765-496-2661, Fax: 765-494-0781

Draft of 4/20/2007

Abstract

Objective: To develop statistical tools that utilize combined initial survival data and post-resuscitation survival data to test the null hypothesis that true, population-wide outcomes following experimental CPR interventions are not different from control.
Method: A new test statistic, d^2, for evaluating Type 1 error is derived from a bivariate, two-dimensional analysis of categorical initial resuscitation and post-resuscitation survival data, which are statistically independent because they are obtained during non-overlapping periods of time. The d^2 test statistic, which is distributed as a chi-squared distribution, is derived from first principles and validated using Monte Carlo methods of computer simulation for thousands of clinical trials.
Results: Under the null hypothesis, the normalized difference in the proportions of patients surviving the initial resuscitation period and the normalized difference in the proportions of such short-term survivors that also survive the post-resuscitation period are jointly distributed in a two-dimensional space as a bivariate standard normal distribution, against which observed intervention and control outcomes can be compared in a test of statistical significance. Typically this two-dimensional approach has greater statistical power to detect true differences, compared to conventional one-dimensional tests. Smaller group sizes (Ns) are usually required to reach statistical significance when both initial survival and post-resuscitation survival are considered together. Such two-dimensional analysis is easily extended to meta-analysis of multiple trials.
Conclusions: A straightforward, easy-to-use bivariate test for Type I errors in statistical inference can be done for resuscitation studies reporting both short-term and long-term survival data. Acceptance of such two-dimensional tests of the null hypothesis, as proposed by Hallstrom, can save time, money, effort, and disappointment in the difficult and sometimes frustrating field of resuscitation research.

Key words: Cardiopulmonary resuscitation (CPR); Clinical trials, Device; Drug Therapy, Meta-analysis, Methodology, Statistical analysis.

---

1. Introduction

In a typical clinical study of a new resuscitation device or method victims of cardiac arrest are randomized to experimental care (the intervention group) and to standard care (the control group). Some measure of outcome, typically survival, is recorded as a categorical end-point. The fractions or percentages of survivors in the intervention group and in the control group are compared for statistical significance. Survival is usually reported both after initial return of spontaneous circulation (ROSC, or short-term survival) and after 24 hours, hospital discharge, or some time after hospital discharge (long-term survival). The numbers of patients surviving long-term tend to be small. Members of the resuscitation research community have debated for years the merits of powering studies for short-term survival versus long-term survival. Ideally one would want to see the virtues of proposed

improvements documented by studies demonstrating improved long-term, neurologically intact survival[1, 2]. No one advocates developing methods that resuscitate hearts but not brains, leaving victims in lingering vegetative states. Yet the awful realities of current long-term survival rates and the unforgiving nature of the binomial distribution dictate that many hundreds, even thousands, of patients must be randomized to achieve a significant long-term end point in a single study.

Pragmatists worry that if all innovations must pass the test of long-term survival, the cost of innovation would become so high that many useful improvements—especially modest, incremental improvements—would never be realized. They retort that since the purpose of CPR is return of spontaneous circulation, and since long-term outcome depends on many confounding factors related to the quality of post-resuscitation intensive care, it is an unrealistic burden to require that all innovations in CPR be tested to the gold standard of long-term, neurologically intact survival. Many fewer patients would be required for statistical significance if the accepted primary end point were ROSC. In turn, clinical research could proceed much faster, ultimately benefiting more people sooner.

The problem with relying on short-term survival only was demonstrated by the experience with high dose epinephrine (some would say any dose of epinephrine), which in animal studies and clinical trials showed increased frequencies of ROSC, but not necessarily long-term survival.[3-10] Further work documented significant toxicity in the form of post-resuscitation myocardial depression and prolonged high peripheral vascular resistance. These effects appear to decrease the chances of surviving the immediate post-resuscitation period, negating any overall long-term benefit and perhaps diminishing the quality of life of those who do survive. The problem in general is that some interventions that increase short-term survival may themselves have long-term toxicity or do lasting harm. Easily imagined examples of such harm include broken ribs, barotrauma to the lungs, myocardial damage, infection, renal failure, hepatotoxicity, or stroke.

This paper is dedicated to the proposition that truly good innovations in CPR will increase both ROSC and post-resuscitation survival and that this combined effect can be verified in more enlightened and efficient statistical tests of the null hypothesis to exclude Type I errors in statistical inference (rejecting the null hypothesis when it is true). In a recent paper utilizing computer simulations Hallstrom[11] has suggested that both short-term and long-term survival are important outcomes and need to be considered jointly. In particular, a bivariate, two-dimensional analysis of survival data can often produce better discrimination of significant results with fewer patients per study group than tests of long-term survival alone. The present paper further develops this concept and presents a simple, direct, analytical approach for two-dimensional tests of joint short-term and long-term survival. Both types of survival data are usually reported in CPR research studies, albeit with smaller numbers for long-term survivors, and so are readily available for analysis. The strategy is to use all hard won clinical data available to test the null hypothesis by explicitly considering survival in two non-overlapping epochs of time—the arrest and resuscitation interval, and the post-resuscitation interval.

2. Methods

2.1 Definitions

Imagine a clinical trial in which N_A patients are randomized to receive standard CPR (controls) and N_B patients are randomized to receive experimental CPR. Let N_{1A} and N_{1B} be the numbers initially resuscitated in the control and intervention groups, according to reasonable criteria for return of spontaneous circulation (ROSC). The corresponding observed proportions of initial survivors are p_{1A} and p_{1B}. Let N_{2A} and N_{2B} be the numbers of patients who survive through the post-resuscitation phase and live long-term according to a reasonable definition, such as neurologically intact hospital discharge. The proportions of long-term survivors, as conventionally reported, are N_{2A}/N_A and N_{2B}/N_B. These values are readily found in the literature as the traditional gold standard end points.

For analysis we divide the trial into two non-overlapping phases in the time domain: the initial resuscitation phase and the post-resuscitation phase. Then we introduce a new outcome measure, which is the proportion of those initially resuscitated that also survive long term. For control group A, $p_{2A} = N_{2A}/N_{1A}$, if $N_{1A} > 0$. For experimental group B, $p_{2B} = N_{2B}/N_{1B}$, if $N_{1B} > 0$. The proportions p_{2A} and p_{2B} are measures of the probability of surviving from the beginning of the post-resuscitation period to the end of the post-resuscitation period. They are statistically independent of proportions p_{1A} and p_{1B} of initial survivors. Proportions p_1 and p_2 represent events in non-overlapping epochs of time. The experiment is like rolling a die twice. Patients have to run certain risks of arrest and CPR itself, and then, if they survive this challenge, run a second set of new and independent risks in the post-resuscitation period.

We consider p_1 and p_2 to be experimental samples of the underlying true probabilities, π_1 and π_2 , of initial resuscitation survival and post-resuscitation survival. For example, if $\pi_1 = \pi_2 = 1/3$ for group A, then the outcome of standard CPR for any single patient could be simulated by rolling a die. Since a normal die has 6 sides, if either one or two spots come up on the first roll, then that would indicate short-term survival. If so, a second roll is taken, and if either one or two spots come up on the second roll, then that would indicate long-term survival. The observed proportions p_{1A} and p_{2A} are the result of repeating this experiment N_A times for group A with true probabilities π_{1A} and π_{2A}. Similarly for group B, the observed proportions p_{1B} and p_{2B} are modeled by N_B paired rolls, with perhaps different probabilities for survival π_{1B} and π_{2B}.

The results will vary according to the underlying binomial probability distributions, which have mean values π and variances $\pi (1-\pi)/N$, where π is the true probability of survival and N is the number of rolls. The clinical trial can be regarded as an experiment designed to measure these true underlying probabilities π_{1A} ,π_{1B} , π_{2A} , and π_{2B} of initial resuscitation survival and post-resuscitation survival in the intervention and control groups by sampling from four different probability distributions. The measured proportions, p_{1A}, ,p_{1B} , p_{2A} , and p_{2B} provide unbiased estimates of the underlying probabilities. These sample proportions have variances $\pi_{1A} (1-\pi_{1A})/N_A$, etc., which are not known exactly at the time

of the experiment. However, unbiased estimates of the variances are provided by $p_{1A}\,(1\text{-}p_{1A})/(N_A - 1)$, etc.[12]

2.2 Two dimensional significance testing

Using these definitions, one may compute the observed initial resuscitation and post-resuscitation survival proportions, p_{1A}, p_{1B}, p_{2A}, and p_{2B} from reported outcome data for control group A and intervention group B. All four proportions are statistically independent. The observed differences in survival proportions between intervention and control groups are

$$\Delta p_1 = p_{1B} - p_{1A} \tag{1a}$$

for initial resuscitation and

$$\Delta p_2 = p_{2B} - p_{2A} \tag{1b}$$

for survival during the post-resuscitation period.

The corresponding variance estimates based upon sample data are

$$\hat{\sigma}_1^2 = \frac{p_{1A}(1-p_{1A})}{N_A - 1} + \frac{p_{1B}(1-p_{1B})}{N_B - 1} \text{ and} \tag{2a}$$

$$\hat{\sigma}_2^2 = \frac{p_{2A}(1-p_{2A})}{N_{1A} - 1} + \frac{p_{2B}(1-p_{2B})}{N_{1B} - 1}, \tag{2b}$$

where $N_{1A} = p_{1A}N_A > 1$, and $N_{1B} = p_{1B}N_B > 1$. Here we invoke the principle that the variance of the sum or difference of two independent random variables is the sum of the variances. The use of N-1 values in the denominators leads to unbiased variance estimates[12].

Under the null hypothesis that $\pi_{1A} = \pi_{1B}$ and $\pi_{2A} = \pi_{2B}$ any apparent differences in outcomes between groups A and B are the result of sampling variation of the binomial distribution. The expected values of Δp_1 and Δp_2 are zero, and the variances are given approximately by (2a) and (2b). Somewhat more stable and accurate estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be obtained for the purpose of null hypothesis testing using pooled estimates for the presumed common survival probabilities in groups A and B during each phase of the study. These pooled estimates are $\bar{p}_1 = (N_{1A} + N_{1B})/(N_A + N_B)$ and $\bar{p}_2 = (N_{2A} + N_{2B})/(N_{1A} + N_{1B})$. Then refined variance estimates can be computed as

$$\hat{\sigma}_1^2 = \frac{\bar{p}_1(1-\bar{p}_1)}{N_A - 1} + \frac{\bar{p}_1(1-\bar{p}_1)}{N_B - 1} \text{ and} \tag{2c}$$

$$\hat{\sigma}_2^2 = \frac{\bar{p}_2(1-\bar{p}_2)}{N_{1A} - 1} + \frac{\bar{p}_2(1-\bar{p}_2)}{N_{1B} - 1}. \tag{2d}$$

In turn, we can define independent normalized differences in survival proportions as

$$z_1 = \frac{\Delta p_1}{\hat{\sigma}_1} \text{ and} \tag{3a}$$

$$z_2 = \frac{\Delta p_2}{\hat{\sigma}_2}. \tag{3b}$$

For N_A and N_B greater than, say, 20 random variables z_1 and z_2 are each distributed as independent standard normal distributions with zero means and approximately unit variances. One can show along the lines of Welch[13] that z_1 and z_2 are each distributed very much like a "Student" t-distribution with a number of degrees of freedom roughly equal to twice the number of survivors in each case. For most practical cases such a Student t-distribution is equivalent to the normal distribution. In turn, to obtain a joint test of the null hypothesis one can compute the test statistic

$$d^2 = z_1^2 + z_2^2, \tag{4}$$

which is distributed approximately as a chi-squared distribution with two degrees of freedom. (In general, if $z_1, z_2, \cdots z_k$ are independent normally distributed random variables, each with zero mean and unit variance, then the random variable $d^2 = z_1^2 + z_2^2 + \cdots + z_k^2$ has, by definition, a chi-squared distribution with k degrees of freedom.)

When the simple variance estimates (2a) and (2b) are used to derive d^2, the actual distribution of (4) will have a slightly longer "tail" than a perfect analytical chi-squared distribution. The tail of the distribution of d^2 more closely approximates the chi-squared distribution when the refined variance estimates (2c) and (2d) are used. As the sizes of the study groups N_A and N_B increase the estimates improve. One objective of the present research is to determine by computer simulation whether for realistic Ns in resuscitation studies, the accuracy of these estimates is sufficient.

One of the pleasing features of this approach is that there is a straightforward graphical interpretation of the test statistic, d^2. It is the square of the straight-line distance between the origin and sample point P = (z_1 , z_2) in the two-dimensional z-space of Figure 1, in which for any study z_1 and z_2 are plotted in a rectangular grid. Here the horizontal axis for z_1 represents the normalized difference in initial survival. The vertical axis z_2 represents the normalized difference in post-resuscitation survival. These two independent axes

represent two non-overlapping phases of resuscitation. Components z_1 and z_2 can be either positive or negative, representing beneficial or harmful effects in each phase. Straight-line distance, d, is the length of the vector between the origin and sample point, $P = (z_1 , z_2)$. Length, d, is always positive or zero. Length, d, is computed from the sum of squares of z_1 and z_2 using the Pythagorean theorem of equation (4).

Under the null hypothesis, denoted H_0 , the probability density distribution for all possible trial outcomes, P, in Figure 1 is binormal, centered about the origin. It has circular contours of constant value in the $z_1 – z_2$ plane. A circular contour at approximately 2.4 units from the origin, such as that labeled H_0 in Figure 1, represents the 95 percent confidence limit of the binormal distribution. If sample point, P, lies at the periphery of this distribution, it is unlikely to be the result of sampling variation only. Test statistic d^2 is a measure of the remoteness of P from the origin. The farther P is from the origin, the smaller is the probability of making a false positive interpretation of the results—that is, a Type I error in statistical inference.

When an alternative hypothesis, H_1, is true, there is a real effect of the intervention upon survival. Then the distribution of sample points will fall farther from the origin in the $z_1 – z_2$ plane. Point P_1 in Figure 1 represents such a result. The distribution of sample points under H_1, also has circular contours but is centered on a point displaced from the origin by a distance proportional to the true treatment effect.

The direction of point, P, from the origin in the $z_1 – z_2$ plane defines whether any significant effect is good, bad, or mixed. A point such as P_1 that is clearly in the right upper quadrant in Figure 1 represents a good effect that improves both short-term survival and post-resuscitation survival from cardiac arrest. A point clearly in the left, lower quadrant represents a harmful effect in both phases of resuscitation, and one located elsewhere represents a mixed effect. Such mixed effects are biologically plausible. A point in the right lower quadrant of the $z_1 – z_2$ plane implies short-term benefit with longer-term toxicity, rather like epinephrine. A point in the left upper quadrant implies short-term toxicity coupled with long-term benefit, for example the action of a cytoprotective drug that also causes hypotension when given as an intravenous bolus (e.g. deferoxamine[14]). Thus it is reasonable to employ a test statistic such as d^2 to evaluate departure from the null hypothesis in all four quadrants. The direction of point, P, from the origin indicates the quality of the effect, and the distance of point, P, from the origin indicates its statistical significance.

Hallstrom[11] has suggested that some points in the right upper quadrant of similar plots that are close to the horizontal (short-term) axis may not represent good outcomes from a cost-effectiveness standpoint. A large increase in short-term survival, accompanied by a very small increase in long-term survival might represent an unreasonable increase in cost, individual suffering, and family suffering for little-long term benefit. Hallstrom's estimate of the unacceptable cost, however, shaves only a small fraction of the area from the bottom of the right upper quadrant, which remains largely intact.

2.3 Monte Carlo methods

To determine if a chi-squared distribution is a reasonable approximation to the actual distribution of d^2 as defined in Equation (4), computer simulations of several million control and experimental resuscitations were implemented in the Visual Basic programming language within a Microsoft Excel spreadsheet. If a random number $0 < x < 1$ was less than a given true probability π_1, then a simulated patient was designated as resuscitated, and in this case if a second random number $0 < x < 1$ was less than a given true probability π_2, then a simulated patient was designated as long-term survivor. For simulations of a true null there was no true difference between intervention and control probabilities: $\pi_1 = \pi_{1A} = \pi_{1B}$, and $\pi_2 = \pi_{2A} = \pi_{2B}$.

To assess agreement between numerical and analytical results, distributions of d^2 test statistics were generated for thousands of simulated clinical trials. Group Ns for each trial ranged from 25 to 200. For each of several trial scenarios the simulated distributions of d^2 test statistics for 50,000 simulated trials was compared with the theoretical chi-squared distribution with two degrees of freedom. To assess agreement in the critical tail regions of the distributions the tail probability (proportion of d^2 values > 6) was calculated using both the numerical results and the analytical chi-squared distribution, for which the tail probability is 0.0498.

2.4 Power of the analysis

Derivation of the power of the analysis is illustrated in Figure 2. Here the distribution of the test statistic, d^2, under the null hypothesis, H_0 , is represented by the thin curve to the left, and the distribution of d^2 under an alternative hypothesis, H_1, of a true positive treatment effect, is represented by the thicker curve to the right. Power may be calculated for the d^2 statistic in the usual way, given an alternative hypothesis $(\Delta\pi_1)^2 + (\Delta\pi_2)^2 > 0$, as well as the associated probability density distribution for $d^2 \,|\, H_1$, and a particular cutoff value, c, for statistical significance of d^2 in a test of the null hypothesis, for example c = 6, which corresponds to 1-tailed $P < 0.05$.

The probability density distribution for $d^2 \,|\, H_1$, is a noncentral chi-squared distribution. The noncentral chi-squared distribution is not as well-known as the ordinary, or central, chi-squared distribution, but it is precisely formulated and can be computed exactly[15]. In particular, if $z_1, z_2, \cdots z_k$ are independent normally distributed random variables with mean values $\mu_1, \mu_2, \cdots \mu_k$ and each with unit variance, then the random variable $d^2 = z_1^2 + z_2^2 + \cdots + z_k^2$ has, by definition, a noncentral chi-squared distribution with k degrees of freedom and noncentrality parameter $\lambda = \mu_1^2 + \mu_2^2 + \cdots + \mu_k^2$. If $\lambda = 0$ then we have an ordinary, central chi-squared distribution, corresponding to the distribution of d^2 under H_0. However, for alternative hypotheses we have $\lambda > 0$. The mean of the noncentral chi-squared distribution is $k + \lambda$, and the variance is $2(k+\lambda)$. Central and noncentral chi-

squared distributions for k = 2 are shown in Figure 2. In the present application for a single research study k is always 2 and $\lambda = \frac{(\Delta\pi_1)^2}{\sigma_1^2} + \frac{(\Delta\pi_2)^2}{\sigma_2^2}$ for true differences in survival proportions $\Delta\pi_1 = \pi_{1B} - \pi_{1A}$ and $\Delta\pi_2 = \pi_{2B} - \pi_{2A}$ with true variances

$$\sigma_1^2 = \frac{\pi_{1A}(1-\pi_{1A})}{N_A} + \frac{\pi_{1B}(1-\pi_{1B})}{N_B} \text{ and } \sigma_2^2 = \frac{\pi_{2A}(1-\pi_{2A})}{\pi_{1A}N_A} + \frac{\pi_{2B}(1-\pi_{2B})}{\pi_{1B}N_B}.$$

Now let f(x) be the known probability density function for the particular noncentral chi-squared distribution[15] of $d^2 \mid H_1$. Let $c = 6.0$ be the cutoff for statistical significance of the central chi-squared distribution with two degrees of freedom, as shown in Figure 2. The power to detect the alternative positive effect is

$$\text{Power} = \int_c^\infty f(x)\,dx . \qquad (5)$$

2.6 Required sample size

Sample size calculations allow investigators to plan study Ns so that there is a high probability of detecting, as statistically significant, a biologically meaningful effect, if it exists. For a particular alternative hypothesis, H_1, for example π_{1A}= 0.2, π_{1B} = 0.3, and π_{2A}= 0.25, π_{2B} = 0.35, one can determine the sample sizes (Ns) required to detect a particular true effect with a particular probability or power. For simplicity the group size for both intervention and control groups is assumed to be the same. By solving equations (1) through (5) together for successive Ns, beginning with a very low value such as 10, a computer program can quickly find the N required for the power to exceed a chosen target value such as 0.90. In such calculations the true, population survival probabilities π_{1A}, etc. and the true population variances $\pi_{1A}(1-\pi_{1A})/N_{1A}$, etc., are used in the place of the sample-based expressions (1) and (2).

3. Results

3.1 Validation of test statistics under H_0

In the theory just presented the test statistic, d^2, should be distributed approximately as a chi-squared distribution with two degrees of freedom under the null hypothesis. To obtain a figure of merit for the goodness of approximation one can compute the tail probabilities of the numerical vs. analytical d^2-distributions. Here we define the tail probability as the area under the probability density function for values of d^2 greater than 6. Histograms of computer simulated results for 50,000 hypothetical clinical studies give a good representation of the actual distribution of d^2. When variance estimates (2a) and (2b) are used to compute d^2, the average tail probability for the d^2 test statistic in 7 computer

simulated scenarios in which the null hypothesis is true was 0.0580 ± 0.0066 SD. This result indicates that actual tail probability is slightly greater under the null hypothesis than that which would be calculated using the chi-square distribution. However, when refined variance estimates (2c) and (2d) are used to compute d^2, the average tail probability for the d^2 test statistic in the same 7 computer simulated scenarios was 0.0495 ± 0.0018 SD (Table 1). The analytical tail probability from the central chi-squared distribution is 0.0498. These results demonstrate that for group sizes typical in resuscitation research the actual distribution of d^2 under the null hypothesis is well approximated by a central chi-squared distribution.

3.2 Sample calculations

Two dimensional significance testing is no more difficult than applying equations (1) through (4) to readily available short-term and long-term survival data from a typical CPR research study that is performed generally according to the Utstein guidelines[16-19]. Consider the hypothetical results for "new" vs. "old" CPR in Table 2. Short-term survival for new CPR is an encouraging 39 percent vs. 25 percent for old CPR. However, the long-term survival is 10 percent for new CPR vs. 11 percent for old CPR. A conventional interpretation of these data might be that although short-term results were encouraging, the gold standard results for long-term survival showed no difference. The conclusion is that the null hypothesis was probably true after all.

Analysis of joint survival leads to a different conclusion. First to compute d^2 it is necessary to find the proportions of short-term survivors who also survive long-term, that is, $p_{2A} = N_{2A}/N_{1A}$ and $p_{2B} = N_{2B}/N_{1B}$. These proportions are different from the conventional proportions of long-term survivors for the entire study, namely N_{2A}/N_A and N_{2B}/N_B. Then it is a simple matter to apply Equations (1) through (4) to compute d^2, as shown in Table 2.

Two-dimensional analysis using the d^2 test shows a significant deviation from the result expected under the null hypothesis, for which we expect $d^2 < 6$. In two-dimensional z-space, the outcome of the trial is represented by a point in the right lower quadrant of Figure 1. With new CPR there is increased short-term survival and decreased survival during the post-resuscitation interval. This result is, if you will, epinephrine-like, showing some lingering toxicity. This two-dimensional statistical inference has consequences for research planning. At first glance one would be tempted to abandon new CPR as altogether ineffective and seek completely new strategies. After analysis of the joint results, however, one would be inspired to modify new CPR to isolate the immediate benefit and minimize the post-resuscitation toxicity.

3.3 Ns needed for significance

Table 3 presents sample (group) sizes for significance in a test of the null hypothesis with 90 percent power over a range of true treatment effects π_{1B} and π_{2B}. Here the group size

for both intervention and control groups is assumed to be the same: $N_A = N_B = N$. The first row and first column of Table 3 show various alternative hypotheses in which survival probabilities for the intervention group are actually greater than those for the control group. For all control groups in Table 3 the probability of initial resuscitation is assumed to be 1/5 and the probability of surviving the post-resuscitation period is assumed to be 1/5 also, so that the probability of long-term survival is (1/5) x (1/5) = 0.04. Various true treatment effects are represented by the column and row values of π_{1B} and π_{2B}, each greater than 0.2. Table entries are the numbers of patients needed to detect a given true effect with 90 percent power. Sample sizes were computed by exhaustive trial-and-error, beginning with a minimal sample size of 10 patients in each group and incrementing the common Ns until 90 percent power is exceeded. Power is computed by numerical integration of the probability density function for the appropriate non-central chi-squared distribution from 6 to infinity. Each value in Table 3 is the group size for a d^2 test with 90 percent power using combined short-term and long-term survival.

For comparison Table 4 gives corresponding sample sizes for a difference of proportions test using long-term survival as the only endpoint. The long-term survival probabilities are $\pi_A = \pi_{1A}\pi_{2A}$ and $\pi_B = \pi_{1B}\pi_{2B}$. Normalized differences in the proportions of survivors are considered to be distributed as standard normal distributions with a mean under the null hypothesis of zero and a mean under the alternative hypothesis of $(\pi_B - \pi_A) \Big/ \sqrt{\pi_A(1-\pi_A)/N + \pi_B(1-\pi_B)/N}$. In most scenarios fewer patients are needed to reject the null hypothesis using the joint d^2 test (Table 3) than when using a single test of long-term survival (Table 4). In some cases half or fewer patients are needed using the joint d^2 test, compared to the one-dimensional z-test using long-term survival alone.

Exceptions occur if the intervention effect on immediate survival is small and the effect during the post-resuscitation interval is somewhat larger. Then a combination of immediate and post-resuscitation outcome has more noise than long-term outcome alone. In this case the joint d^2 test requires larger Ns than a simple z-test of long-term survival. For most other cases however, the joint test has greater power.

4. Discussion

Putative improvements in CPR that increase ROSC but leave survivors in lingering vegetative states who never regain a semblance of health are not to be desired and would substantially increase health care costs and family suffering. For this reason it is important not to claim a good outcome of a randomized clinical trial of a new CPR technique without improvement in long-term survival. However, the numbers of patients required for direct statistical tests of long-term survival data are often prohibitively large, owing to the nature of the binomial distribution and the statistical sampling thereof, especially when control survival probability is low. It is possible, however, to use both short-term ROSC and post-resuscitation survival together to test the null hypothesis with greater power. If this is done in a way that ensures statistical independence of the initial resuscitation and post-resuscitation results, then a simple and straightforward test using the chi-squared probability distribution can be conducted.

One merely computes the test statistic $d^2 = z_1^2 + z_2^2$ for a study and compares the result to the number 6. This approach is easy to present visually and tends to demystify the process of two-dimensional, bivariate analysis, rendering it accessible to physician-scientists who have had an introductory course in statistics. It is also a stronger and more conservative test of the null hypothesis because it is able to detect biologically important mixed results, such as increased probability of ROSC coupled with subsequent decreased probability of surviving the post-resuscitation period. Such subtle, harmful effects in the post-resuscitation phase might be missed with analysis of long-term survival alone. On the other hand the d^2 test is also more sensitive to true departures from the null hypothesis for most plausible trial scenarios.

The method can also be extended to a meta-analysis of multiple studies of similar interventions in a very straightforward way. For k studies of essentially the same intervention that report both long term and short term data one merely adds together the z_1^2 and z_2^2 components in equation (3) from the various studies to get an expanded chi-squared with 2k degrees of freedom. If a study lacks long-term data, a degree of freedom can be subtracted. The result is distributed as a higher degree chi-squared under the null hypothesis, which will have a cutoff value for rejecting the null hypothesis > 6, depending on the degrees of freedom.

It is important to emphasize that proportion p_2 in the forgoing discussion is not equal to conventional long-term survival (namely, the number of long term survivors divided by the number randomized for each group) but instead the proportion of patients surviving short-term that also survive long-term. This is a measure of the probability of surviving the post-resuscitation period only. In this case statistical independence is assured.

Use of d^2 test statistics that are distributed as noncentral chi-squared distributions when there is a given true effect of the intervention allows for direct estimation of the power of a study and also the N's required to demonstrate an expected treatment effect. If Table 3 does not suffice for planning of studies, a computer program embedded in an Excel spreadsheet for evaluating the noncentral chi-squared distribution is available electronically from the author upon request or can be written *de novo* without much trouble.

The traditional discussion of outcomes in CPR research has been framed in terms of "either-or". We should use either short-term or long-term survival as the primary endpoint in resuscitation studies. Why not use both to extract the maximum amount of information from hard-won clinical data? This approach may provide a way to satisfy both the short-term and long-term camps debating the proper outcome measures of CPR studies and also to save time, money, effort, and disappointment in clinical resuscitation research. Alfred Hallstrom's idea that short-term and long-term survival data need to be considered jointly is a good one.

References

1. 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; 95:2213-2239.
2. Cummins RO, Chamberlain DA, Abramson NS, et al. Recommended guidelines for uniform reporting of data from out-of-hospital cardiac arrest: the Utstein Style. A statement for health professionals from a task force of the American Heart Association, the European Resuscitation Council, the Heart and Stroke Foundation of Canada, and the Australian Resuscitation Council. Circulation 1991; 84:960-75.
3. Brown CG, Martin DR, Pepe PE, et al. A comparison of standard-dose and high-dose epinephrine in cardiac arrest outside the hospital. New Engl J Med 1992; 327:1051-1055.
4. Ditchey RV. High-dose epinephrine does not improve the balance between myocardial oxygen supply and demand during cardiopulmonary resuscitation in dogs, JACC 3, 1984.
5. Pearson JW, Redding JS. The role of epinephrine in cardiac resuscitation, Anesth Analg 42, 1963.
6. Stell IG, Herbert PC, Weitzman BN, et al. High dose epinephrine in adult cardiac arrest. New Engl J Med 1992; 327:1045-1050.
7. 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.
8. Marwick TH, Case C, Siskind V, Woodhouse SP. Adverse effect of early high-dose adrenaline on outcome of ventricular fibrillation. Lancet 1988; 2:66-8.
9. Babbs CF, Berg RA, Kette F, et al. Use of pressors in the treatment of cardiac arrest. Ann Emerg Med 2001; 37:S152-62.
10. Cao L, Weil MH, Sun S, Tang W. Vasopressor agents for cardiopulmonary resuscitation. J Cardiovasc Pharmacol Ther 2003; 8:115-21.
11. Hallstrom AP. What is the appropriate outcome for studies of treatments for out-of-hospital cardiac arrest? Resuscitation 2006; 71:194-203.
12. Babbs CF. Simplified meta-analysis of clinical trials in resuscitation. Resuscitation 2003; 57:245-55.
13. Welch BL. The significance of the difference between two means when the population variances are unequal. Biometrika 1937; 29:350-362.
14. Kompala SD, Babbs CF, Blaho KE. Effect of deferoxamine on late deaths following cardiopulmonary resuscitation in rats. Annals of Emergency Medicine 1986; 15:405-407.
15. Winer BJ. Statisitcal Principles in Experimental Design. New York: McGraw-Hill, 1971:907 pages.
16. Patrick A, Rankin N. The in-hospital Utstein style: use in reporting outcome from cardiac arrest in Middlemore Hospital 1995-1996. Resuscitation 1998; 36:91-4.
17. 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.

18. Cummins RO. The Utstein style for uniform reporting of data from out-of-hospital cardiac arrest. Ann Emerg Med 1993; 22:37-40.
19. 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.

Tables

Table 1. Computer simulated tail probabilities for the d^2 test statistic calculated using variance estimates (2c) and (2d) in runs of 50,000 simulated clinical trials

N	π=0.25	π=0.5
200	0.0490	0.0484
100	0.0527	0.0509
50	0.0488	0.0475
25		0.0490

Common group size is $N = N_A = N_B$. Common survival probability is $\pi = \pi_{1A} = \pi_{1B} = \pi_{2A} = \pi_{2B}$. The mean tail probability is 0.0495 ± 0.0018 SD. Analytical tail probability for the chi-square distribution is 0.0498.

Table 2. Hypothetical outcome data for a study of "new" CPR (group B) vs. "old" CPR (group A) with $N_A = N_B = 100$ patients in each group

	ROSC N_1	Discharge survivors N_2	p_1	p_2	p_3
Group A	25	11	0.25	0.44	0.11
Group B	39	10	0.39	0.256	0.10

Proportions: $p_1 = N_1/100$, $p_2 = N_1/N_2$, $p_3 = N_2/100$.

Statistics using simple variance estimates (2a) and (2b): $z_1 = 2.14$, $z_2 = -1.48$, $d^2 = 6.77$.

Statistics using refined variance estimates (2c) and (2d): $z_1 = 2.11$, $z_2 = -1.50$, $d^2 = 6.71$.

Table 3. Sample size for significance in a test of H_0 when H_1 is true using the d^2 test of joint short-term and long-term survival

	π_{2B}					
π_{1B}	0.25	0.3	0.35	0.4	0.45	0.5
0.25	1001	982	641	436	310	229
0.3	454	388	315	251	200	159
0.35	218	201	178	154	132	111
0.4	128	122	113	102	91	80
0.45	84	81	77	72	66	60
0.5	59	58	55	53	49	46

Table 4. Sample size for significance in a test of H_0 when H_1 is true using a difference of proportion test for long-term survivors

	π_{2B}					
π_{1B}	0.25	0.3	0.35	0.4	0.45	0.5
0.25	1001	925	551	375	277	215
0.3	925	506	330	237	181	145
0.35	551	330	226	167	131	106
0.4	375	237	167	127	100	82
0.45	277	181	131	100	80	66
0.5	215	145	106	82	66	54

Figures

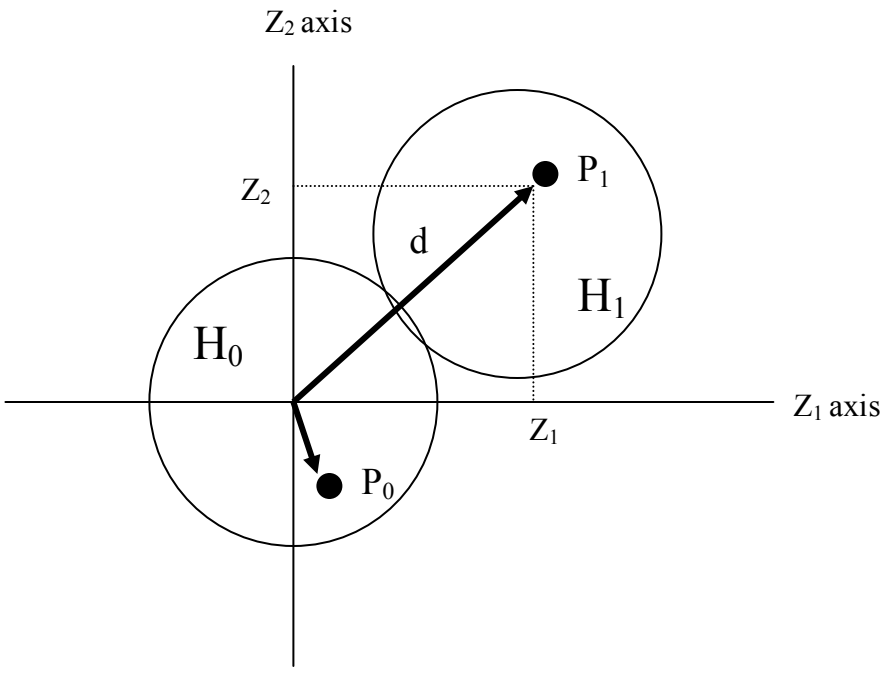

Figure 1. Graphical representation of two-dimensional survival analysis. Component test statistics z_1 and z_2 represent normalized differences in immediate survival on the horizontal axis and in post-resuscitation survival on the vertical axis. Joint test statistic, d^2, is the square of the straight-line distance from the origin of the $z_1 – z_2$ plane to a sample point such as P_1, representing the results of a particular study. Under the null hypothesis, H_0, sample points such as P_0 will tend to cluster about the origin (contour H_0). Under an alternative hypothesis, H_1, sample points such as P_1 cluster farther from the origin (contour H_1).

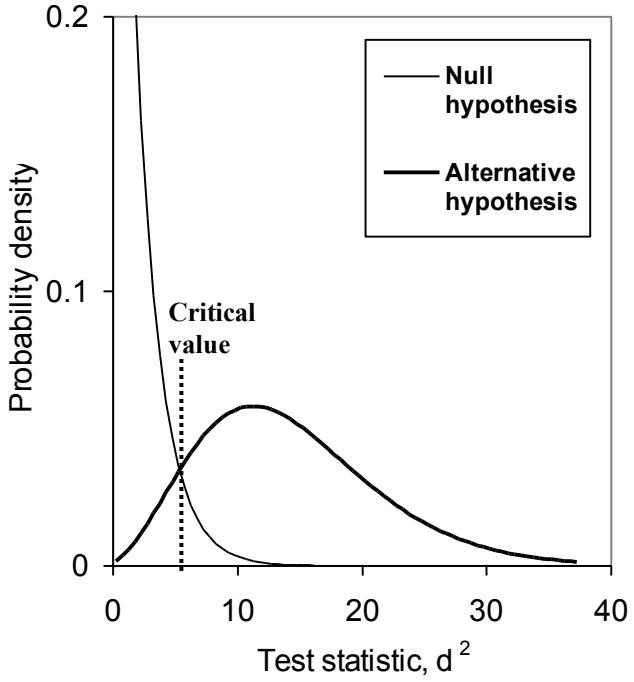

Figure 2. Calculation of power from probability density distributions for the null hypothesis, thin curve, and for an alternative hypothesis, thick curve. The critical value for significance is x = 6.0. The area under the thick curve to the right of the critical value is the power.