EFFECTIVENESS ESTIMATES FOR SMALL CASE-CONTROL STUDIES WITH DICHOTOMOUS OUTCOMES

by

Donald I. Promish
68 Richardson Street
Burlington, Vermont
05401-5026

23 December 2012

Abstract

This article combines Bernoulli's and Bayes's theorems to produce a tool for analyzing case-control studies whose outcomes are dichotomous. The study cohorts can range upwards in size starting at 2. Tabulated examples demonstrate that the law of large numbers (which ensures that very large samples are highly representative of the populations from which they are drawn) applies only to large numbers, and not to small numbers. The article offers a simple gauge of the evidentiary strength of a case-control study. Results include demonstration analyses of real case-control statistics in the fields of posttraumatic stress disorder (PTSD), breast cancer screening and terrorism.

Keywords

Forensic, small studies, effectiveness, likelihood ratio, probability, evidence.

Introduction

This article arises from my work in forensic identification (Promish & Lester, (1999); Promish, (2008a,b)) and from the fact Dror & Rosenthal (2008) use the effect size indicator $r_{equivalent}$ developed by Rosenthal & Rubin (2003). Dror & Rosenthal (2008) attempt to evaluate fingerprint examiners; however, their data do not meet the requirements for valid use of $r_{equivalent}$.

If, 40 years on, psychological researchers [e.g., Dror & Rosenthal (2008)] are still flouting the warnings of Tversky & Kahneman (1971), possibly to the detriment of the forensic science community; and if, as recently as 2003, psychological researchers [e.g., Rosenthal & Rubin (2003)] are still unsatisfied with their attempts at a "simple effect size indicator", I think there is no harm in offering my viewpoint on the problem of the small trial.

The fingerprint examiner outcomes in Dror & Rosenthal (2008) consist of only two alternatives, "match" and "no match". For example, fingerprint expert C's reliability study data, taken from Dror & Rosenthal (2008), with the original fingerprint examination results playing the roles of "case" and "control" and the retest results acting as the outcomes, appear in Table 1. It is obvious that the retest "outcomes" are dichotomous and thus not normally distributed. The data also is sparse.

The reader can perform a simple analog of the "test-retest" study used in Dror&Rosenthal (2008), by tossing a coin twice on one day and twice again on the following day. Letting H stand for heads and T stand for tails, the possible paired toss outcomes are (H T), (T H), (H H) and (T T). It is doubtful whether one could conclude, from any of these "studies", whether the coin is fair or biased, and if biased, whether toward H or T. Suppose that, a day after producing (H H), for example, the same coin yields (T T). These events do not necessarily mean that the coin was biased toward heads the day before, and changed its bias overnight. The coin could be perfectly fair; yet, under the extremely limited observation of 4 tosses, it will seem both biased and unreliable, as the following demonstration shows.

Here is a sequence of 50 outcomes of "tosses" of a computer-simulated fair coin:
H H T **H** H T **T** T **H** H T T T T H T H H T H H H H T H T T H T H T T T **T** T H H **H** H T T T T T H T H H H T.
The computer displayed "H" for a "toss" if its random number generator produced a value, v, in the range ($0.5 \leq v \leq 1$); it displayed "T" for v in the range ($0 \leq v < 0.5$). There are, it should be noted, 24 "H"s and 26 "T"s. Of these 50 tosses, only 47 can begin a 4-toss sub-sequence. Of these 47 tosses (reading from left to right), 5 bold-face, underscored outcomes (**H** or **T**) begin 4-toss sub-sequences (i.e., **H**HTT or **T**THH) which wrongly suggest an unreliable, biased coin. Because each toss of a fair coin has 2 equally-probable outcomes, a 4-toss sequence has ($2\times2\times2\times2=$) 16 equally-probable outcomes. Only 2 of those outcomes, HHTT and TTHH, wrongly suggest unreliable bias.

So, in the long run, a fair coin will produce such sub-sequences (100% × 2/16=) 12.5% of the time. [The proportion for the short sequence above is (100% × 5/47 =) 10.6%.] The 4-toss observer thus runs a 12.5% risk of mislabelling a fair coin as unreliable and biased. As some of the other 4-toss sub-sequences above suggest, there are other pitfalls, just as likely, awaiting this observer.

Rosenthal & Rubin (2003) define $r_{equivalent}$ as equal to "the sample point-biserial correlation between the treatment indicator and an exactly normally distributed outcome in a two-treatment experiment ...". They emphasize that the more the actual outcome distribution differs from exact normality, "the less relevant is the approximation using $r_{equivalent}$." As shown above, the sparse, dichotomous statistics of Dror & Rosenthal (2008) fail even to approximate a normal distribution.

Rosenthal & Rubin (2003) conclude their paper in the hope that, in view of the limitations of $r_{equivalent}$, a "highly sophisticated" alternative can be found for it. My primary aim here is to provide a simple analytic tool which can do what $r_{equivalent}$ (for example) cannot: analyze small studies whose outcomes are dichotomous. Secondarily, I aim to show that no analytic tool can be expected to produce convincing results from sparse data.

Derivation of the Bernoulli-Bayes method

The method that I propose relies on the binomial theorem of Jakob Bernoulli, and Thomas Bayes's theorem "on the doctrine of chances".

Consider a small study of a disease treatment whose effect is unknown (except, perhaps, anecdotally). One part of the trial cohort consists of the treated subjects, while the other part, the control subjects, receive no treatment. Suppose, also, that there are two possible outcomes of the study, which could be (improvement/no improvement), (cure/no cure) or (survival/death), depending on the nature of the disease.

The question to be answered by our small "treatment" vs "control" study is, "How does the likelihood of improvement or cure or survival, given treatment [abbreviated P(ics|gt)], compare with the likelihood of improvement or cure or survival denied treatment [P(ics|dt)]?" The study being small, the outcome data is not only dichotomous; it is also sparse.

This article derives, as the central measure of the study's results, the mean of the collection of all possible ratios of the form P(ics|gt)/P(ics|dt), called likelihood ratios. It expresses the quantitative "spread" of this collection by means of their standard error.

Under each condition, treatment and control, the subjects' outcomes can be modeled as a sequence of Bernoulli trials, one Bernoulli trial per subject. Under each condition (treatment/control) all the Bernoulli trials are assumed to have the same single-trial likelihood of success (i.e., improvement/cure/survival). Before the study, we know neither of these two constant values; hence we must assign probabilities to all the possible values (ranging from 0 to 1) of each. Our first step, then, is to develop, for each condition, the probability distribution of its single-trial likelihood of success. We use Bayes's theorem in order to do this.

Bayes's theorem, when applied to the outcomes of a series of Bernoulli trials, yields the probability distribution of the single-trial likelihood of success, as follows. For $1 \leq j \leq 20$, I define p_j as the midpoint of each of the 20 likelihood intervals (1.00,0.95), (0.95,0.90), (0.90,0.85), ... (0.05,0.00).

Bayes's theorem then calculates the probability of each hypothesis H_j, which says: "The single-trial likelihood of success underlying this series of Bernoulli trials is p_j".

[As will be seen, the 20-interval partition suffices not only to expose the uncertainty that results from a very small sample, but also to get useful results from samples up to nearly 600 subjects in size. The 20-interval partition yields 400-element double sums for the mean likelihood ratio and its standard deviation in Equations (4) and (5), respectively. One could use a finer partition; however, the limit of the method's validity using 20 intervals appears well within the sample size region where the method becomes unnecessary.]

The evidence, D, to be considered here is that k successes out of n trials have occurred under either the treatment condition or the control condition, where n is the number of subjects and k is the number of improvements/cures/survivals, and $0 \leq k \leq n$. The probability of the evidence D (given the hypothesis H_j regarding the single-trial likelihood of success p_j) is, by the formula for the Bernoulli distribution,

$$P(D \mid H_j) = \left(\frac{n!}{k!(n-k)!} \right) p_j^{\,k} \left(1 - p_j\right)^{n-k} \quad . \tag{1}$$

By substitution, the Bernoulli formula (1) appears below in Bayes's theorem (2), for the probability of a particular hypothesis H_j (in the numerator), given the evidence D and its likelihoods for all possible hypotheses H_i (in the denominator) :

$$P(H_j \mid D) = \frac{P_0(H_j)P(D \mid H_j)}{\sum_{i=1}^{i=20} P_0(H_i)P(D \mid H_i)} . \tag{2}$$

If we take each of the 20 *a priori* probabilities, $P_0(H_{j(\text{or } i)})$, to be equal [i.e., $P_0(H_{j(\text{or } i)}) = 1/20$, for $1 \leq j \text{ (or i)} \leq 20$] on the grounds that the only information we have is the observed study evidence D, then Bayes's theorem reduces to

$$P(H_j \mid D) = \frac{P(D \mid H_j)}{\sum_{i=1}^{i=20} P(D \mid H_i)} . \tag{3}$$

Now, first suppose our small "treatment" vs "control" study consists of a "treatment" series of only 2 Bernoulli trials resulting in 1 success.

[Here is a worked example of the computations, using Equations (1) and (3), that produce the right-hand column of Table 2 for the "treatment" series. The single-trial probability of success may have one of the following 20 values, as shown in the left-hand column of the table: {0.975, 0.925, 0.875, 0.825, 0.775, 0.725, 0.675, 0.625, 0.575, 0.525, 0.475, 0.425, 0.375, 0.325, 0.275, 0.225, 0.175, 0.125, 0.075, 0.025}. Equation (1) assigns, in order, the following probabilities of 1 success in 2 trials to each of these values: {0.05, 0.14, 0.22, 0.29, 0.35, 0.40, 0.44, 0.47, 0.49, 0.50, 0.50, 0.49, 0.47, 0.44, 0.40, 0.35, 0.29, 0.22, 0.14, 0.05}. The sum of these probabilities is 6.68; this is the denominator of Equation (3). Corresponding, for this example, to a single-trial probability of success of 0.725, Equation (1) gives 0.40 as the probability of 1 success in 2 trials; this is the numerator of Equation (3). Then the likelihood that 0.725 is the single-trial probability of success for the treatment series turns out, according to Equation (3), to be (0.40/6.68 =) 0.0597.]

As Table 2 shows, one cannot infer from this series that the single-trial likelihood of "treatment" success is exactly 0.5; it is actually more apt to be between 0.75 and 1 (0.0522 + 0.0433 + 0.0328 + 0.0208 + 0.0073 = 0.1564) than it is to be between 0.45 and 0.55 (0.0747 +0.0747 = 0.1494).

Second, imagine our small study has a "control" series of only 2 Bernoulli trials that result in no success. This does not imply that the single-trial likelihood of "control" success is exactly zero. As Table 3 (which is the "control" counterpart of Table 2) shows, it is more apt to be between 0.40 and 0.60 (0.0496+0.0414+0.0339+0.0271 = 0.1520) than it is to be between 0 and 0.05 (0.1427).

Because we are comparing two conditions, giving and denying of treatment, we now replace the single index, "j", by two indices. Using "g" to index success likelihoods **g**iven treatment and "d" to index success likelihoods **d**enied treatment, we see that no single "g,d" pair taken from Tables 2 and 3 will answer the question conclusively. Our first step in the analysis is to find all the likelihood ratios, for $1 \leq (g \text{ or } d) \leq 20$, of the kind

(success likelihood "g" given treatment)/(success likelihood "d" given no treatment).

Then we weight each likelihood ratio by the joint probability

[probability of (success likelihood "g" given treatment)] ×

[probability of (success likelihood "d" given no treatment)].

Each term in this product is calculated according to Equation 3.

Next, we add up the weighted likelihood ratios to get a mean likelihood ratio. Finally, we calculate the standard error of the mean likelihood ratio.

For instance, in Table 2 ("treatment") we find a value of 0.725 for the likelihood of success given treatment, and its probability 0.0597; and in Table 3 ("control") we find a value of 0.125 for the likelihood of success given no treatment, and its probability 0.1149. The weighted ratio of these two likelihoods is, then,

$$(0.725/0.125) \times [0.0597 \times 0.1149], \text{ or about } 0.040.$$

By taking the weighted double sum over <u>all</u> likelihoods of "treatment" success and "control" success (20 values for each), we obtain the mean likelihood ratio, $\overline{\rho}$:

$$\overline{\rho} = \sum_{g=1}^{g=20} \sum_{d=1}^{d=20} \left(p_g / p_d\right) \times \left[P(p_g) \times P(p_d)\right] \quad . \qquad (4)$$

For the small study we have been discussing, $\overline{\rho} = 5.2$.

The standard error, SE, of the mean likelihood ratio can easily be calculated by the formula for the difference between the mean of the squares, and the square of the mean, of the likelihood ratios; it is the positive square root of the variance:

$$SE = \sqrt{\sum_{g=1}^{g=20} \sum_{d=1}^{d=20} \left(p_g / p_d\right)^2 \times \left[P\left(p_g\right) \times P\left(p_d\right)\right] - \left\{\overline{\rho}\right\}^2} \quad . \qquad (5)$$

Again, for the study under discussion, SE = 7.3 .

There is one value of the likelihood ratio against which our results can usefully be compared. That value is 1. "$\rho = 1$" means that the results we obtained under "treatment" (or, more generally, "case") conditions are just as likely to occur under "control" conditions; and that, therefore, that we cannot tell at all whether the "treatment" has had any effect. By analogy with the null hypothesis, H_{null}, which posits that our "case" observations are the result purely of chance, we can define the "null likelihood ratio", ρ_{null}, thus: $\rho_{null} \equiv 1$.

If the mean likelihood ratio is either greater than or less than ρ_{null}, then our "case-control" results tend to confirm that "case" outcomes differ from "control" ones. However, since the likelihood ratio is a random variable, we have to express the difference between the mean likelihood ratio and ρ_{null} in terms of the "spread" of ρ, expressed here by the SE. In this article, I use the absolute value of $((\overline{\rho} - 1)/SE)$ as the gauge, γ, of the evidentiary value of the study:

$$\gamma \equiv |((\overline{\rho} - 1)/SE)| \quad . \qquad (6)$$

The results $\overline{\rho} = 5.2$ and SE = 7.3 from our exemplary small study, above, tell us that $\gamma = [((5.2 - 1)/7.3) =] 0.58$. We see that the mean likelihood ratio for our small study is well within one SE of the value $\rho = 1$. Our results are suggestive but not conclusive; perhaps not even persuasive.

Results and discussion

Table 4 displays the mean likelihood ratios for notional studies of survival under alternative conditions of treatment and control. I took studies 1 through 7 from Rosenthal and Rubin (2003); I then extended the list with notional studies 8 through 10.

As an example of interpretation, study 6 tells us that, on average, a patient is about 14 times as likely to survive with treatment as without treatment. However, the gauge $\gamma = [(13.8 - 1)/12.9 =] 0.99$ warns us not to take this apparently impressive result at face value.

The notional study cohorts in Table 4 range in size from 2 to 200. In all of the studies, the mean likelihood ratios for survival are greater than 1, suggesting that, to a greater or lesser extent, the treatment works. However, until cohort sizes increase toward 40 (study 8), for which $\gamma = [(28.8 - 1)/13.1 =] 2.1$, I would not be comfortable claiming that the treatment is as effective as one would like. I might infer from the statistics of, say, study 6, that a larger study is justified; or I could simply wait until more data turns up.

Passing from the notional to the factual, I now present applications of the proposed method to three contemporary real-world topics: (1) posttraumatic stress disorder; (2) breast cancer screening; and (3) terrorism. These applications suggest that the Bernoulli-Bayes method not only works well on any problem whose data format is "case-control" with dichotomous outcomes, but also that its results approach "large study" results as sample sizes rise into the hundreds. This, of course, is as it should be.

Posttraumatic stress disorder

Consider the claim that the United States government should not take responsibility for the care of combat veterans exhibiting posttraumatic stress disorder (PTSD), because "they would have developed PTSD even if they had never seen combat".

Gilbertson, McFarlane et al. (2010) studied identical-twin pairs of which only one member saw combat, and as a consequence rebut these claims. They write, "Combat veterans with PTSD demonstrated significantly higher scores on ... psychometric measures of psychopathology than their own combat-unexposed cotwins (and than combat veterans without PTSD and their cotwins). ... These results support the conclusion that the majority of psychiatric symptoms reported by combat veterans with PTSD would not have been present were it not for their exposure to traumatic events."

In order to apply the Bernoulli-Bayes method, I adopt Gilbertson, McFarlane et al. (2010)'s assumption that "a combat veteran's non-combat-exposed identical twin is a valid surrogate for what the veteran would have become absent the experience of combat ...".

Gilbertson, McFarlane et al. (2010) describe twin-pairs whose combat-exposed members were rated, on average, at 3.8 (SE = 2.7) on a Combat Severity Scale (CSS), and did not have PTSD; there were 54 such pairs. In Bernoullian terms, these twin-pairs represent 54 trials yielding 0 successes.

Gilbertson, McFarlane et al. (2010) also describe twin-pairs whose combat-exposed members were rated, on average, at 7.7 (SE = 2.5) on the CSS, and did have PTSD; there were 50 such pairs. In Bernoullian terms, these twin-pairs represent 50 trials yielding 50 successes.

The Bernoulli-Bayes mean likelihood ratio for the twin-pair sets described above is 37.3; its standard error is 6.1. Hence, $\gamma = [(37.3 - 1)/6.1 =] 5.9$, and the fact that the mean likelihood ratio is approximately 6 times the standard error strongly suggests that a combat veteran whose severity of experience (per the CSS) is about 8 would be roughly 37 times as likely to develop PTSD as would a veteran whose average severity of experience is about 4 (or, indeed, their unexposed twins).

(Where, on the Combat Severity Scale, the transition from non-PTSD to PTSD induction occurs is one question; another is whether the location and rate of the transition depend on monozygotic twinhood.)

Breast cancer screening

Chronix Biomedical (2010) announced a "DNA blood [test that detects] breast ... cancer with 92% sensitivity and 100% specificity", on the basis of samples comprising 178 women with early stage breast cancer and 200 healthy controls. The claim of 100% specificity is particularly suspect because 100% of 200 is still only 200; it clearly calls for analysis by the Bernoulli-Bayes method proposed here.

Using the Chronix Biomedical (2010) data, I find that, for calculating the chance that a patient has breast cancer, a positive Chronix assay has a mean likelihood ratio of 39.0 (SE 0.1); and a negative Chronix assay has a mean likelihood ratio of 0.0833 (SE 0.0172).

In contrast, Chronix Biomedical (2010) mention that data from a large study of U.S. mammography screening programs reported an overall sensitivity of 75% and a specificity of 92.3% . Straightforward calculation yields, for a positive mammogram, the likelihood ratio (0.75/(1 - 0.923) =) 9.7; for a negative mammogram, the likelihood ratio is ((1 - 0.75)/0.923 =) 0.271.

Both Chronix mean likelihood ratios are further away from the value $\rho = 1.0...$, than the mammography ratios are; the positive assay ratio by ((39.0 - 1)/0.1 =) 380 SEs, the negative assay ratio by about ((1 - 0.0833)/ 0.0172 ≅) 54 SEs. In other words, γ for the positive assay is 380, and γ for the negative assay is approximately 54. Thus, the Chronix assay offers credibly stronger evidence than mammography does of either the presence or absence of breast cancer.

Terrorism

On 12 September 2010, The New York Times (NYT) Sunday Magazine feature "Idea Lab" [Berreby (2010)] discussed Gambetta & Hertog (2009). The writer, David Berreby, inferred from this item that, "Had [violent Islamist] groups reflected the working-age populations of their countries, engineers would have made up about 3.5 percent of the membership. Instead, nearly 20 percent of the militants had engineering degrees." This comparison amounts to a likelihood ratio, without taking sample size (404 militants) into account, of (20/3.5=) 5.7. That is, the writer concludes that engineers are nearly 6 times as likely to be found in violent Islamic groups as they are in the general working population they come from.

Soon after I read the Berreby article, I used its numbers to make a Bernoulli-Bayes estimate (which does consider sample size) of the likelihood ratio. I found that engineers are 7 to 9 times as likely to be found in a violent Islamist group as they are in the general working population they come from. My estimate is close to Berreby's while showing that even a sample size of 404 yields a spread of values.

More recently, I acquired Gambetta & Hertog (2007), which is the working paper predecessor of Gambetta & Hertog (2009). From this source, I constructed a case-control study comparing engineer membership in violent Islamist groups [Gambetta & Hertog (2007), Table 9] with engineer membership in non-violent Islamist groups [ibid., Table 10]. Limiting the study to groups in Southeast Asia, the Middle East and North Africa, I found that about 80 of a sample of 254 violent Islamists (~ 31.5%) were engineers; while only 59 of 585 non-violent Islamists (10%) were engineers. Disregarding sample size, it appears that engineers are about 3 times as likely to be found in a violent group as in a non-violent one of similar geographic/ideological makeup.

The Bernoulli-Bayes method (which does take sample size into account) yields a mean likelihood ratio of 3.0, with, in addition, a standard error of 0.8. Thus, because both the violent and the non-violent sample sizes are well over 100, and also because $\gamma = ((3.0 - 1)/0.8 =) 2.5$, the Bernoulli-Bayes results agree with those of the preceding paragraph, while adding quantitative meaning to the word “about” found there.

The least that can be expected of this article is further refutation, if such is needed, of what Tversky and Kahneman (1971) have called the "law of small numbers". The law of large numbers ensures that very large samples are highly representative of the populations from which they are drawn. The "law of small numbers" is the mistaken belief that the law of large numbers applies to small numbers as well. The Introduction of this article should have dispelled it.

The "Derivation ..." section of the article uses the artifice of a small case-control trial with dichotomous outcomes in order to demonstrate, step-by-step, how Bernoullian and Bayesian concepts, when combined, lead to the easily-computed, simple measures of "center" (mean likelihood ratio) and "spread" (standard error of the likelihood ratio) for evaluating both the strength and the credibility of the results of a small trial.

The "Results" part of the present section demonstrate not only that the proposed method works for small trials, but that it produces answers, for moderately large trials, that are compatible with straightforward likelihood calculations that can safely assume zero spread in their results. It also shows that the method is situationally versatile because of its simple requirements for valid use.

References

Berreby, B., (2010). *Engineering Terror*. The New York Times Sunday Magazine, 12 September 2010.

Chronix Biomedical, (2010). *Data Presented at ASCO show Chronix Biomedical's DNA blood tests detect breast and prostate cancer with 92% sensitivity and 100% specificity*. www.chronixbiomedical.com, Chicago, IL and San Jose, CA, June 7, 2010.

Dror, I. & Rosenthal, R. (2008). *Meta-analytically quantifying the reliability and biasability of forensic experts.* Journal of Forensic Sciences, 53, 4, pp. 900-903.

Gambetta, D., & Hertog, S. (2007). *Engineers of jihad*. Sociology working paper number 2007-10. University of Oxford Department of Sociology. [http://www.nuff.ox.ac.uk/users/gambetta/engineers%20of%20jihad.pdf]

Gambetta, D., & Hertog, S. (2009). *Engineers of jihad*. European Journal of Sociology, 50, 02, pp. 201 - 230. doi: 10.1017/S0003975609990129

Gilbertson, M.W., McFarlane, A.C. et al. (2010). *Is trauma a causal agent of psychopathologic symptoms in posttraumatic stress disorder? Findings from identical twins discordant for combat exposure.* Journal of Clinical Psychiatry, 71, 10, pp. 1324 - 1330.

Promish, D.I. & Lester, D. (1999). *Classifying serial killers*. Forensic Science International, 105 (1999) pp. 155 - 159.

Promish, D.I. (2008a). *Monte Carlo Bayesian identification using STR profiles.* Available from the National Criminal Justice Reference Service (www.ncjrs.gov); posted with NCJ number 221192.

Promish, D.I. (2008b). *Monte Carlo Bayesian identification using SNP profiles.* Available from the National Criminal Justice Reference Service (www.ncjrs.gov); posted with NCJ number 224106.

Rosenthal, R. & Rubin, D. (2003). $r_{equivalent}$: *A simple effect size indicator.* Psychological Methods, 8, 4, pp. 492-496.

Tversky, A. & Kahneman, D. (1971). *Belief in the law of small numbers.* Psychological Bulletin 2, pp. 105-110.

Table 1. Fingerprint expert C's reliability study data, from Dror&Rosenthal (2008). The original fingerprint examination results play the role of "case" and "control"; the retest results act as the outcomes.

Retest	Original test	
	Match	No match
Match	3	0
No match	1	4

Table 2. Probability distribution of the single-trial likelihood of success in a series of 2 Bernoulli "treatment" trials resulting in 1 success. The probabilities were calculated at the midpoints of the 20 likelihood intervals [1.00,0.95], [0.95,0.90], [0.90,0.85], ... [0.05,0.00] .

Single-trial likelihood of success, p	Probability of p given 1 success in 2 trials
0.975	0.0073
0.925	0.0208
0.875	0.0328
0.825	0.0433
0.775	0.0522
0.725	0.0597
0.675	0.0657
0.625	0.0702
0.575	0.0732
0.525	0.0747
0.475	0.0747
0.425	0.0732
0.375	0.0702
0.325	0.0657
0.275	0.0597
0.225	0.0522
0.175	0.0433
0.125	0.0328
0.075	0.0208
0.025	0.0073

Table 3. Probability distribution of the single-trial likelihood of success in a series of 2 Bernoulli "control" trials resulting in 0 success (i.e., 2 failures). The probabilities were calculated at the midpoints of the 20 likelihood intervals [1.00,0.95], [0.95,0.90], [0.90,0.85], ... [0.05,0.00] .

Single-trial likelihood of success, p	Probability of p given 0 success in 2 trials
0.975	0.0001
0.925	0.0008
0.875	0.0023
0.825	0.0046
0.775	0.0076
0.725	0.0114
0.675	0.0159
0.625	0.0211
0.575	0.0271
0.525	0.0339
0.475	0.0414
0.425	0.0496
0.375	0.0586
0.325	0.0684
0.275	0.0789
0.225	0.0902
0.175	0.1022
0.125	0.1149
0.075	0.1284
0.025	0.1427

Table 4. Several notional studies showing how treatment effectiveness estimates improve with increasing sample size.

Study	Conditions (number of subjects)	Outcomes		Analysis	
		Survivals	Deaths	Mean likelihood ratio for survival (Treatment//Control)	Standard error of mean likelihood ratio for survival
1	Treatment (1)	1	0	5.3	8.0
	Control (1)	0	1		
2	Treatment (2)	2	0	5.9	8.7
	Control (1)	0	1		
3	Treatment (2)	2	0	7.8	10.0
	Control (2)	0	2		
4	Treatment (3)	3	0	8.3	10.4
	Control (2)	0	2		
5	Treatment (3)	3	0	10.0	11.3
	Control (3)	0	3		
6	Treatment (5)	5	0	13.8	12.9
	Control (5)	0	5		
7	Treatment (10)	10	0	20.7	14.2
	Control (10)	0	10		
8	Treatment (20)	20	0	28.8	13.1
	Control (20)	0	20		
9	Treatment (50)	50	0	37.0	6.8
	Control (50)	0	50		
10	Treatment (100)	100	0	38.9	1.9
	Control (100)	0	100		