

© 2026 Mikel Aickin – CC BY-NC 4.0

Better, Worse, Same

The Analysis and Interpretation of Success/Failure Probabilities

Mikel Aickin
Minneapolis, Minnesota, USA
mikelaickin@gmail.com
Version 16 Oct 2026

Abstract. A success probability provides a very simple summary of how well some group is doing, and so it is natural to use comparisons of success probabilities across groups as an attempt to quantify which groups are doing better. But all such conventional comparisons are made from the viewpoint of the observer, not from the outlook of someone trying to choose which group to join. Choice of a particular group typically imposes a cost of some kind, and so the decision must be based on the trade-off between the cost and the probabilistic benefit. The purpose here is to define the probability of benefit and to provide a method to analyze it, with particular attention to choosing medical treatments.

Introduction

One of the most standard of all inferential procedures involves the comparison of the "success" probabilities in two groups, with the usual binomial assumptions, along with the presumption of independence of the two groups in which the probabilities are estimated. A conventional comparison of the probabilities is expressed in terms of their difference, or their ratio, or in epidemiology their odds ratio. But when the outcomes really can be interpreted as successes or failures, there is an alternative comparison that sometimes makes more sense. This is especially true in clinical science. The purpose of this note is

to explain this approach, to provide the inferential tools, and to discuss its applicability.

Probabilities of Better, Worse, and Same

For notation, let p be the probability of a success in the first group, with n opportunities, and q be the probability of a success in the second group, with m opportunities. If we were to draw a pair of measurement instances at random, one from each group, then the probability that the first would be a *better* result than the second is p(1-q). The probability that the first would be *worse* than the second is (1-p)q, and the probability they would be *the same* is pq + (1-p)(1-q). The usual estimates of p and q would be the fractions of successes in each group, and standard computations give the variances of the estimates of the probabilities of better, worse, and same, as shown in Table 1 (see the Appendix).

Interpretation	Probability	Variance
Better	p(1-q)	$p(1-q)\left(\frac{1-p}{n}+\frac{q}{m}\right)-r$
Worse	q(1-p)	$q(1-p)\left(\frac{p}{n}+\frac{1-q}{m}\right)-r$
Same	pq + (1-p)(1-q)	$\frac{p(1-p)}{n}+\frac{q(1-q)}{m}-4r$
Both Success	pq	$pq\left(\frac{1-p}{n}+\frac{1-q}{m}\right)-r$
Both Failure	(1-p)(1-q)	$(1-p)(1-q)\left(\frac{p}{n}+\frac{q}{m}\right)-r$
Better - Worse	p-q	$\frac{p(1-p)}{n}+\frac{q(1-q)}{m}$

Table 1. Definitions of comparison probabilities and their variances.

$$r = p(1-p)q(1-q)\left(1-\left(1-\frac{1}{n}\right)\left(1-\frac{1}{m}\right)\right)$$

The results in Table 1 can be used for any statistical method based on the Normal approximation to the binomial, but for the purpose of this note I will use support functions, as defined in Aickin (2025). Briefly, the support function is defined over all possible values of a parameter, and displays for each value the support that the data give to it. Support can be conventionally interpreted as a (two-sided) p-value, or as one minus the maximum confidence one can have to rule it out using a confidence interval. Zero support means that the data give no support to the parameter value, and a support of 1 means the data give maximal

support. The potential value of support functions was pointed out by Birnbaum (1961), and advocated for use in epidemiology by Poole (1987).

Support functions will be computed here for the three parameters, probability of "better", probability of "worse", and probability of "same". The interpretation of the results in terms of Better/Worse/Same is not equivalent to conventional comparisons of p and q. In a clinical application, a success is a good medical outcome in some well-defined way, and the first group is treated while the second receives usual care, or a placebo. Conceptually, the issue can be symbolized as "good medical outcome ⟵ treatment". The intent of the analysis presented here is to inform the patient that if he/she chooses treatment, then pr(Better) is the probability that they will succeed when they would have failed on usual care, and pr(Worse) is the probability that they will fail when they would have succeeded on usual care. And pr(Same) is the probability that their choice would have no effect.

Although it is not necessary to the discussion, this interpretation is consistent with the causal "counterfactual" viewpoint. In this conceptualization, before the treatment is accepted or refused the patient is regarded to have their own individual response profile, telling which outcome they will experience as a function of which therapy they choose. "Causal" in this approach means that what actually happens is simply determined by the therapeutic choice applied to the patient's response profile.

Extension of the clinical metaphor to more general situations can vary. In the case of a social intervention, with the intent of gaining some benefit for those who receive it, the analogy seems fairly strong. But when group membership is determined in some other way, an argument may be required to clarify the underlying interpretation.

In the following sections I present three applications, two explicitly medical and one socio-cultural.

No COVID symptomology ⟵ hydroxycholorquine

This was a randomized clinical trial, in which HCQ was administered to about half of a sample of 1360 health care workers who were presumably at high risk of COVID-19 infection (Naggle et al. 2023). A negative outcome was clinical or presumptive COVID symptoms within 30 days of commencement of the trial, and so a success was its opposite. The issue can therefore be phrased in success/failure terms.

The inferential analysis is given in Table 2, and the support functions for pr(Better) and pr(Worse) are in Fig. 1. Note that pr(Same) is quite large, due to the fact that COVID symptoms were relatively rare in the studied groups. That is, in a high proportion of cases, it made no difference whether HCQ was administered or not, and this was because most had no symptoms. In Fig. 1, the minimum of the two support functions is a sub-support function for a supposed common value of p and q, and its maximum (0.274) is the support for the claim that they are equal, and then the maximally supported common value is 0.065. There is, therefore modest evidence that HCQ is ineffective.

Note, however, that if one were to accept that pr(Better) = pr(Worse) on this basis, then there is another value of pr(Better) which has the same support (0.274), it is pr(Better) = 0.085, and in the light of the results it should then be equally acceptable. The investigators concluded no effect based on a confirmation of a null hypothesis test, but they did not consider what difference between p and q had the same support as p=q, and therefore they made no assessment of the potential benefit of HCQ. The analysis here would agree with the investigators, since the difference between 0.085 and 0.065 is so small, however (a) the outcome was not COVID, but a suggestion of COVID; (b) the longer-range effect on clinically diagnosed COVID was not investigated, suggesting that abandonment of HCQ might be premature.

Interpretation	**Probability**	**SDE**
Better	0.074	0.0097
Worse	0.055	0.0084
Same	0.871	0.0119
Both Success	0.866	0.0128
Both Failure	0.005	0.0009
Better - Worse	0.018	0.0138

Table 2. Comparative probability estimates for the hydroxychloroquine study.

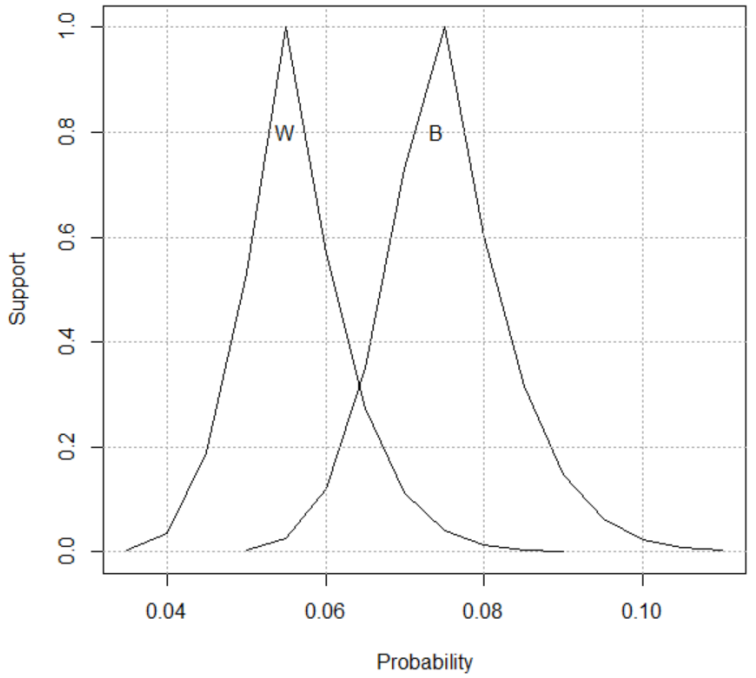


Fig. 1. Support functions for pr(Better) and pr(Worse) in the hydroxychloroquine study. Their minimum is the sub-support function for a supposed common value of the two parameters.

No MS Events ← Continued Therapy

The second example was also a randomized trial, this time to determine whether continuing therapy to prevent multiple sclerosis symptoms was effective (Corboy et al. 2023). 259 MS patients were studied over two years, with a success being the absence of MS relapse or appearance of new symptoms. The numerical results are in Table 3 and the support functions in Fig. 2. The latter suggest that the better-supported parameter values of pr(Better) are about 0.07 larger that pr(Worse), indicating a supported benefit of continuing therapy. The support for a common value is 0.107, occurring at probability values of 0.08 and 0.17. Thus, if one were to accept that there were no detectable difference (both equal 0.08) then one would similarly have to accept that pr(Better) = 0.17. That is, a no-benefit decision would be as well-supported as a decision for a considerable benefit. The conclusion is, therefore, that the study may not have provided a sufficiently precise answer to the question.

The investigators concluded that they had not detected a benefit of continued therapy, which was odd because it conflicted with the confidence interval they gave for the probability difference.

Interpretation	Probability	SDE
Better	0.116	0.0274
Worse	0.041	0.0165
Same	0.842	0.0295
Both Success	0.837	0.0318
Both Failure	0.006	0.0027
Better - Worse	0.075	0.0342

Table 3. Comparative probability estimates for the multiple sclerosis study.

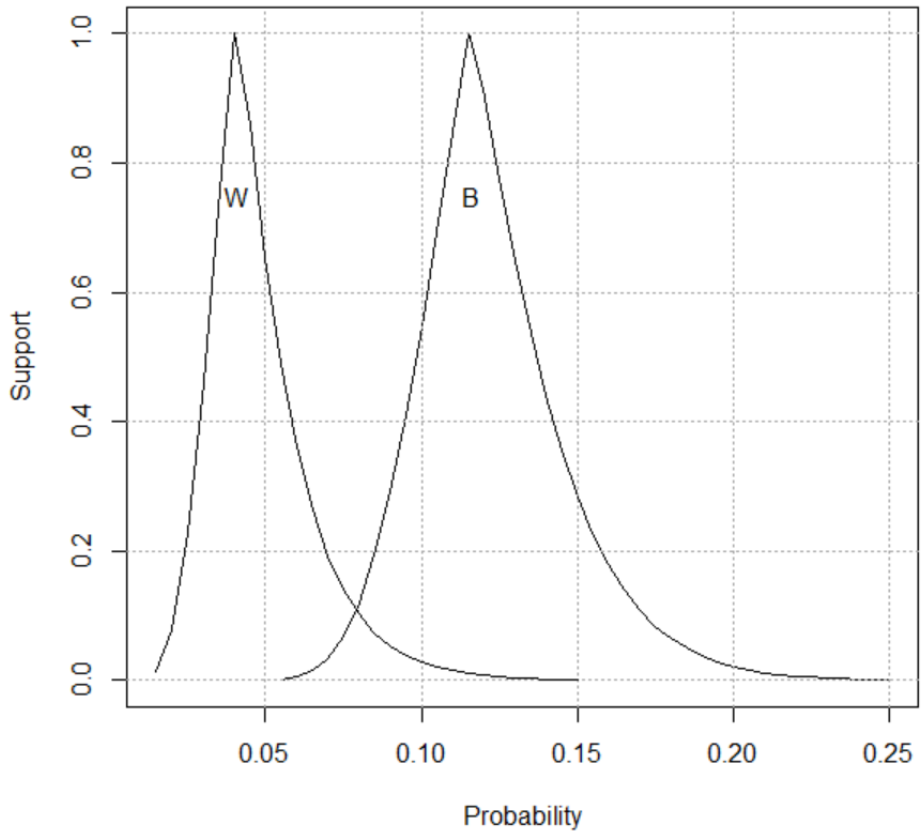


Fig. 2. Support functions for pr(Better) and pr(Worse) in the MS study.

Lifeboat ← Female Selection

One of the most-analyzed tables in statistics courses involves death or survivorship among the passengers on the Titanic, which sank in the North Atlantic in 1912. The appreciation of the iceberg-collision damage was rather slow, which meant that lifeboats were loaded and set afloat while it was still unclear whether the ship was in danger of sinking. Since there were not enough lifeboats for all the passengers, whether one got a place in a boat was an important determinant of survival. From passenger narratives the story emerged that initially there was a "women and children first" ethic in deciding who went into the boats. This created the "survival ← female" issue, which is indeed suggested by the form in which the table of counts is usually presented.

One objection to such an analysis, raised by at least one causal theorist, was that gender is not a choice, and so it is inappropriate to introduce causal ideas into the Titanic survival story. For us this illustrates that moving from clinical decision making to wider sociological situations might require some additional interpretation. In this case, one can focus on the process of *deciding who was to get the available lifeboat seats*, and here the *choice* was indeed one of gender – which gender to give life-boat preference to. Using "survival" as a surrogate for getting a lifeboat seat, the model proposed in the title of this section seems appropriate for the type of analysis we are considering. The issue then becomes possible determinants of how seats were allocated, and this has in turn focused on the classes of the passengers.

	Probability (SDE)		
	1st Class	2nd Class	3rd Class
Better	0.656 (0.0357)	0.789 {0.0377)	0.386 {0.0335)
Worse	0.009 {0.0046)	0.012 (0.0043)	0.088 {0.0112)
Same	0.335 (0.0338)	0.200 (0.0337)	0.527 (0.0263)
Both Success	0.317 (0.0347)	0.072 (0.0186)	0.075 (0.0101)
Both Failure	0.019 (0.0093)	0.128 (0.0331)	0.452 (0.0338)
Better - Worse	0.647 (0.0380)	0.777 (0.0418)	0.298 (0.0424)

Table 4. Comparative probability estimates for the Titanic sample. SDE = standard deviation of the estimate.

The support functions in Fig. 3 make it apparent that passenger class modified the gender selection process. One might have hypothesized that the tendency to benefit females would have been in the same order as class, but

instead it was stronger for the second class passengers than for the first class passengers. Although one would naturally turn to passenger narratives for an explanation, there is a hint in the pr(Both Successes) and pr(Both Failures) lines in Table 4. The first class passengers had a tendency to be Both Successes while the third class passengers had a tendency to be Both Failures. These probabilities were smaller and closer together in the second class passengers. A hypothesis might then be that the decisions in the two extreme classes were more clear-cut, to give seats in first class and deny them in third class, but among second class decisions the choice was more difficult, and this was where the favoring of females was expressed. Stated differently, a crew member confronted with first class passengers wanting seats did not have to make a male-female distinction, because the tendency was to give either a seat, while among third class passengers it was to give neither a seat. Discrimination occurs most readily where decisions are the most difficult.

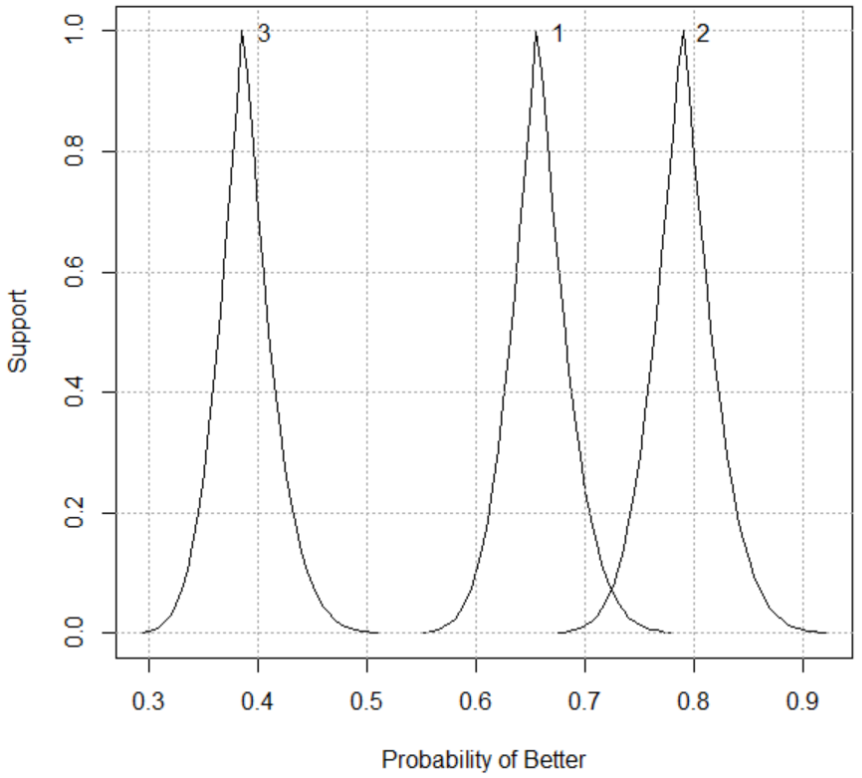


Fig. 3. Support functions for pr(Better) for the three passenger classes on the Titanic.

Comments

Choice of medical treatment based on success probabilities was most prominently promoted by P.C.A. Louis, in Paris in the 1830's. His approach was considered unconventional because it was based on how treatments did in groups of patients, instead of the then-dominant method of making specific judgments about continued therapy based on how the individual patient responded to the initial course of therapy. Poisson invented the first confidence intervals, for a binomial proportion and a difference between binomial proportions, in his 1837 book on using probabilities for judgments. Although it was emphasized by Gavarret in the first biostatistics book (1840), it made no headway until the expansive development of biostatistics in Britain, initiated by Karl Pearson in the last 5 years of the 19th century.

The use of statistics in medical research has almost exclusively concentrated on comparisons of parameters, considered as characteristics of populations or sub-populations. In the process, the notion that medicine involves the treatment of individual patients was absorbed by the competing concept that it is primarily about treating groups of patients. The perspective offered here is that clinical medicine should focus on probabilities of better, worse, or same, at the level of the individual patient. In some ways this goes back to the 19th century debate, whether medicine was art or science, with a solution that is a proper combination of the two.

The basic strategy of direct comparison of pairs of patients is capable of very considerable generalization, including specific subgroup results as well as direct control of confounding factors (Aickin 2023). The method presented here is only the simplest special case.

Summary

Analysis based on the better/worse/same trichotomy is of a different character than conventional approaches based on metrics that compare the probabilities p and q directly. As illustrated in the examples, the metaphor fits most easily in the context of clinical medicine, where the idea was born. But the extension to social intervention studies may not be very difficult, and as the Titanic example shows, close interpretation of the actual social process being observed in a non-interventional setting may also narrow the gap.

As the probability tables suggest, the tri-outcome analysis offers more insight into what might have actually been taking place in context. This feature is enhanced by using support functions, in place of conventional hypothesis tests or confidence intervals. This is especially true in the case of hypothesis tests, where decision-making has been only half-heartedly incorporated into basic inference, instead of being seen as a process that follows and depends on inference.

The basic approach of considering hypothetical comparisons as a means of explaining chance-affected processes is not new, but it has usually been exiled to the fringe of statistical methods. The version given in this note is particularly

simple, being based on an imagined comparison of two measurement instances drawn from two groups. There is, however, a far more advanced version which involves selecting analytic strata and matching on potential confounding factors (Aickin 2023).

References

Aickin M. Support functions: Reframing conventional statistical methods. *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5705144* (2025: SSRN website)

Aickin M. Matched comparison pairs. *www.lecturelessnotes.org*. click on *Scientific Inference: Matched Comparison Pairs*. 2023

Birnbaum A. Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association* 1961:56(294); 246-249

Corboy JR, Fox RJ, Kister I, Cutter GR, Morgan CJ, Seale R, Engebretson E, Gustafson T, Miller AE. Risk of new disease activity in patients with multiple sclerosis who continue or discontinue disease-modifying therapies. *The Lancet Neurology* 2023;22:568-577

Naggie S, Milstone A, Castro M, Collins SP, Lakshmi S, Anderson DJ, Cahuayme-Zuniga L, Turner KB, Cohen LW, Currier J, Farulo E, Friedland A, Garg J, George A, Mulder H, Olson R, O'Brien EC, Rothman RL, Shenkman E, Shostak J, Hernandez AF. Hydroxycholorquine for pre-exposure prophylais of COVID-19 in health care workers. *International Journal of Infectious Diseases* 2023;129:40-48

Poole C. Beyond the confidence interval. *American Journal of Public Health* 1987;77(2):195-199

Appendix: Variance Computations

For any independent chance variables x and y which have finite variances,

$$\mathrm{var}[xy] = \mathrm{E}[x^2y^2] - \mathrm{E}[x]^2\mathrm{E}[y]^2 = \mathrm{var}[x]\mathrm{var}[y] - \mathrm{E}[x]^2\mathrm{var}[y] - \mathrm{E}[y^2]\mathrm{var}[x]$$

Writing $\hat{p}$ and $\hat{q}$ for independent binomial fractions estimating p and q, based on n and m opportunities respectively, and writing p*=1-p, the above equation becomes

$$\mathrm{var}[\hat{p}\hat{q}] = \frac{pp^*qq^*}{nm} - \frac{p^2qq^*}{m} - \frac{q^2pp^*}{n} = pq\left(\frac{p^*q^*}{nm} - \frac{pq^*}{m} - \frac{qp^*}{n}\right) =$$

$$= pq\left(\frac{q^*}{m} + \frac{p^*}{n}\right) - pp^*qq^*\left(1 - (1 - \frac{1}{n})(1 - \frac{1}{m})\right)$$

The first two results in Table 1 are obtained by substituting q* for q or p* for p.

For the probability of "same" we need

$$\mathrm{cov}[\hat{p}\hat{q}, \hat{p}^*\hat{q}^*] = \mathrm{E}[\hat{p}\hat{q}\hat{p}^*\hat{q}^*] - \mathrm{E}[\hat{p}\hat{q}]\mathrm{E}[\hat{p}^*\hat{q}^*] = \mathrm{E}[\hat{p}\hat{p}^*]\mathrm{E}[\hat{q}\hat{q}^*] - pp^*qq^* =$$

$$= \left(p - \left(\frac{pp^*}{n} + p^2\right)\right)\left(q - \left(\frac{qq^*}{m} + q^2\right)\right) - pp^*qq^* = -pp^*qq^*\left(1 - (1 - \frac{1}{n})(1 - \frac{1}{m})\right)$$

Therefore,

$$\mathrm{var}[\hat{p}\hat{q} + \hat{p}^*\hat{q}^*] = \mathrm{var}[\hat{p}\hat{q}] + \mathrm{var}[\hat{p}^*\hat{q}^*] - 2pp^*qq^*\left(1 - (1 - \frac{1}{n})(1 - \frac{1}{m})\right) =$$

$$= pq\left(\frac{p^*}{n} + \frac{q^*}{m}\right) + p^*q^*\left(\frac{p}{n} + \frac{q}{m}\right) - 4\,pp^*qq^*\left(1 - \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{m}\right)\right) =$$

$$= \frac{pp^*}{n} + \frac{qq^*}{m} - 4\,pp^*qq^*\left(1 - \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{m}\right)\right)$$

CC BY-NC 4.0

Creative Commons Attribution-NonCommercial 4.0 International

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only.