

Evidence-Based Medicine

A Primer on Medical Statistics – P Values and Confidence Intervals

Evette Weil, B.A. (OT2)

Introduction

This article is a continuation of the evidence-based medicine series that began in the May 1999 issue of the UTMJ. Evidence-based medicine requires that the physician have a fundamental understanding of statistical concepts in order to critically interpret the published literature. This and future articles in the next two issues of the UTMJ are intended to serve as an introductory primer on statistics relevant to the clinician. This article starts with the most basic measures used in clinical epidemiology, p values and confidence intervals, measures that are likely to be familiar to most readers. Future articles will examine in detail other statistical concepts, such as odds ratios, risk reduction, number needed to treat, correlation, and regression.

Case Study

Fibrate therapy of low HDL levels reduces major cardiovascular events

This study investigated the use of a fibrate, gemfibrozil, in the secondary prevention of coronary heart disease in men with low levels of high density lipoprotein (HDL) cholesterol. In a double-blind trial, 2531 men with coronary artery disease and a primary lipid disorder of low HDL levels were randomized to gemfibrozil therapy or placebo. Patients had normal or low triglyceride and low density lipoprotein (LDL) levels. At one year, HDL cholesterol was increased by 6%, triglycerides were decreased by 31%, and LDL cholesterol did not change. A primary event, defined as myocardial infarction or death from coronary causes, occurred in 17.3% of the gemfibrozil group and in 21.7% of the placebo group. The relative risk reduction was 22% (95% confidence interval: 7 to 35%; $p=0.006$). There was no significant difference in rates of coronary revasculariza-

tion, hospitalization for unstable angina, death from any cause, and cancer. This study suggests that fibrate therapy may have a role in the treatment of patients with coronary artery disease and low levels of HDL.

Source: Rubins *et al.* (1999). *NEJM*. 341:410-418

1. Significance

1.1 The Importance of Chance

Statistical analysis of a clinical trial assumes that there is a true treatment effect, but that the trial can only estimate that effect. There is always the possibility that an apparently large effect found in a particular sample is due to chance. This notion of chance is integral to all statistical measures.

The conventional model for explaining the idea of chance is a coin-toss experiment. We can intuitively understand that on any given coin toss there is a 50% chance that the result will be heads. However, we can imagine a scenario of a fair coin being tossed 10 times will result in 7 heads and 3 tails. The divergence from the 50:50 ratio is due to chance. Likewise, in a clinical trial testing the efficacy of a new drug, the true benefit of the drug may be a 50% reduction in mortality, but the amount measured may be 30% or 70%. This potential variation complicates our ability to judge the validity and applicability of the results.

Let us return to our coin toss experiment. We may accept that in an unbiased test, 7 heads may result from 10 coin tosses. However, if we tossed the coin 10 times, and heads came up 9 times, it would be more difficult to ascribe the results simply to chance. Similarly, we would be less ready to accept that 70 out of 100 or 700 out of 1000 coin tosses could be due to chance.

With these results, we would suspect that there were some external factor favouring the heads side of the coin. These two examples demonstrate that the likelihood of a given result being real, as opposed to chance, is dependent on the magnitude of the outcome and the number of independent observations (sample size).

1.2 P Values

The variation between real and random findings may be represented statistically with p values. A p value is the probability that a given result occurred by chance. Using our coin toss experiment, we can calculate with a statistical formula (binomial expansion) the chance of obtaining at least 7 heads in 10 coin tosses. This number is the p value of the experiment. For 10 coin tosses, the probability of seeing 7 or more heads (or the same arrangement with tails) is calculated to be 17%. In other words, in 17 out of 100 experiments, we would expect to see at least 7 heads or 7 tails; so $p=0.17$.

Now imagine, instead of a coin toss, a clinical experiment where 7 out of 10 patients experienced a favourable outcome, and a statistical test generates a p value of 0.17. This means that there is a 17% probability that the observed result is merely due to chance. The question now becomes one of clinical rather than statistical judgement: Do we consider a result with such a p value to be significant?

For most scientific literature, the arbitrary line drawn between a significant and non-significant finding is $p < 0.05$. This implies that the medical profession considers a result noteworthy if the odds are less than 1 in 20 that the result is due to chance. Highly significant results are $p < 0.01$ or odds of 1 in 100 that the result is due to chance.¹

We can illustrate the use of significance with the clinical trial described above. In this trial, researchers compared a drug, gemfibrozil, to placebo for the secondary prevention of coronary events in patients with low HDL cholesterol.² The primary outcome for the study was nonfatal myocardial infarction or death from coronary causes. The study found that 17.3% of patients given gemfibrozil had a primary event versus 21.7% of patients given placebo. This represents a 22% reduction in risk. More importantly, this reduction has a p value of 0.006, and is, therefore, highly significant. (Note: More comprehensive explanations of risk reduction will be given in the next article in this series).

The gemfibrozil study had as a secondary outcome death from any cause. In this case, 15.7% of patients given gemfibrozil died compared to 17.4% of those given placebo. This represents an 11% reduction in risk. However, the associated p value is only 0.23, and is considered non-significant. How are we to understand these figures? We cannot interpret this as proving that gemfibrozil had no effect on overall mortality. We can only say that the study was not large enough – or, in epidemiological jargon, did not have enough power – to determine whether or not gemfibrozil had an effect on death from any cause.

1.3 Multiple Outcomes

The fact that the gemfibrozil study tested multiple outcomes (one primary and 9 secondary) brings up a prevalent problem in the medical literature. When testing for multiple outcomes, there is an increased likelihood that a particular result is due to chance. If we have 10 outcomes, each with an individual p value of 0.05 (i.e. 95% chance that the result is valid and not merely due to chance), the probability that all 10 of them simultaneously are valid is $(0.95)^{10} = 0.60$. In other words, there is a 40% probability that at least one of the 10 results is due to chance despite the individual p values of 0.05. A more comprehensive explanation of this phenomenon is given by Guyatt *et al*.³

From a practical point of view, this means that with multiple endpoints we have to reduce our threshold for significance to compensate for the fact that more of our results can be ascribed to chance. One simple strategy for managing this problem – the one used in gemfibrozil study – is to distinguish between a single primary outcome and multiple secondary outcomes. The standard p value threshold of 0.05 can be applied to the primary outcome. Therefore, the lower incidence of the primary outcome of myocardial infarction or death from coronary heart disease, $p = 0.006$, remains significant. For the multiple secondary outcomes, the threshold p value is divided by the number of measured outcomes. In this case, the new threshold becomes 0.05 divided by 9 or 0.006. Now we can even make a stronger case that the lower incidence of the secondary outcome of death from any cause, $p = 0.23$, is not significant.

2. Confidence Intervals

2.1 An Additional Way to Report Results

The concept that measured results may be due either to a real effect or to chance can also be expressed statistically with confidence intervals (CI). In reports of clinical trials, a single result (point estimate) is often given with both a p value and a confidence interval. A confidence interval relates the point estimate of a given parameter to the true population parameter. In other words, a 95% CI means that there is a 95% probability that the true value lies in the given range, assuming a normal distribution. Strictly speaking, a 95% confidence interval means that if the study were repeated 100 times and CIs constructed in the same way, 95% of these CIs would contain the true value. The first definition is more comprehensible, but statisticians dislike it, because it implies that the true value is changing. CI's gives a range of included and excluded values on which to base clinical decisions in the face of uncertainty.

In the case study, the point estimate for the reduction in non-fatal myocardial infarction or death due to coronary heart disease is 22% for gemfibrozil versus placebo. The authors also report that the 95% confidence interval for this measure is 7 to 35% risk reduction. This means that we can be 95% confident that the *true* risk reduction would lie in this range. Since the lowest limit of this range, +7, is still above zero, the authors can be

95% confident that even their most conservative estimate shows a benefit from gemfibrozil.

In contrast, the confidence interval for relative risk reduction of death from any cause ranges from -8 to 27%. The negative value means that at the lower limit, gemfibrozil may actually increase the all-cause mortality risk. How can we interpret these figures? Our point estimate, 11% risk reduction with gemfibrozil, is our best single estimate of the true outcome. All of the other values in the confidence interval distribute normally around the point estimate. We can conclude that this study showed a trend toward risk reduction, but it is not definitive, because the lower limit of our confidence interval includes the possibility that the drug actually increases all-cause mortality.

2.2 The value of confidence intervals

Confidence intervals are particularly useful because they allow the reader to determine whether a study was large enough. As the sample size of a study increases, the confidence interval becomes narrower, since, as was noted above, with more "tries" fewer results can be attributed to chance. So with a small study, we are likely to obtain a large confidence interval with greater likelihood that the two limits would fall on either side of a threshold value.

The threshold value depends on the nature of the trial and on clinical judgement. A larger study will always give a narrower CI, but is it worthwhile to do a larger trial? Consider a highly toxic drug used to treat some non-life threatening condition. Suppose the first trial showed a 4% improvement in symptoms with a CI of -2% to 10%. The upper confidence limit suggests that the real benefit is unlikely to be greater than a 10% improvement, and we might say that such a benefit "isn't worth it" since the drug is so toxic. Clearly we are now describing clinical significance rather than statistical significance. In contrast, in the gemfibrozil study, the upper confidence limit suggested that the benefit might be as high as a 27% relative reduction in all-cause mortality, something that may well be clinically important. Researchers and clinicians need to decide where to place the threshold; deciding, for example, that only a risk reduction greater than 5% would merit a change in practice.

In general, the rules for interpreting confidence intervals are as follows: 1) For a trial that appears to be positive, if the lower limit is above the threshold of what would be considered a clinically important effect, the study is positive and definitive; 2) for a trial that appears positive, but the lower limit is below the threshold, the study is likely to be positive, but is not definitive; likewise, 3) for a trial that appears negative, but the upper limit of the confidence interval is above the threshold, the study is likely to be negative, but is not definitive; and finally, 4) for a trial that appears to be negative and the upper limit is below the threshold, the results can be viewed as definitively negative (Table 1).

Table 1
Rules for Interpreting Confidence Intervals

Consider a situation where a 10% improvement in symptoms is required in order to justify the use of the therapy.

Result: % change in symptoms (95% CI)	Classification	Interpretation
27% (15%, 42%)	Positive and definitive	Positive because the CI does not include zero and definitive because the lower CI is greater than our clinical threshold of 10%
14% (2%, 28%)	Positive but not definitive	Not definitive because the lower CI is less than our clinical threshold of 10%
-3% (-12%, 15%)	Negative but not definitive	Negative because the CI crosses zero, but not definitive because the upper CI includes the possibility of clinically significant benefit
-1% (-6%, 8%)	Negative and definitive	Negative because the CI crosses zero and definitive because upper CI is below our clinical threshold of 10%

Conclusion

The concepts of *p* values and confidence intervals highlight the point that results reported from clinical trials are only estimates of the *true* effect of the intervention. *P* values and confidence intervals allow the physician to read an article and make a reasonable decision after regarding the validity of the results. It is this judgement that prompts the physician to consider altering his or her practices, and, therefore, represents the fundamental step in practicing evidence-based medicine.

Acknowledgments

The author would like to thank Dr. Jan Hux at the Institute for Clinical Evaluative Sciences for her help in reviewing this article.

References

- Greenhalgh T. (1997). How to read a paper: Statistics for the non-statistician. *BMJ*. 315: 422-425.
- Rubins HB, Robins SJ, Collins D, *et al* (1999). Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol. *NEJM*. 341(6): 410-418.
- Guyatt G, Jaeschke R, Heddle N, *et al* (1995). Basic statistics for clinicians. *CMAJ*. 152: 27-32.