Statistical Analysis, Reporting, and Interpretation of Glaucoma Surgical Trials: A Brief Practice Guide

William J Feuer

Division of Biostatistics, Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, Miami, FL USA

INTRODUCTION

Most clinicians and clinician scientists will have had an introduction to statistics, either in medical school coursework or through their own study. However, the practice of statistics in medicine is a burgeoning discipline (witness the growth of the respected journal *Statistics in Medicine* since it was introduced in the early 1980s) and difficult to keep abreast of. The aim of this article will be to summarize some key concepts and techniques in biostatistics with particular reference to clinical trials, not from first principles (this is an interval level variable; that is an ordinal variable), but rather as a guide to analyses and practices commonly employed in ophthalmology journals. The format of this article is to introduce a variety of these techniques and identify some of the salient features concerning their interpretation. This summary is particularly relevant to surgical trials where treatments are compared with respect to incidence rates of treatment failure, rather than, say, medical trials that may compare average differences in intraocular pressure (IOP) lowering. The emphasis is on heuristic exposition rather than technical precision.

Comparison of Treatments: Kaplan-Meier Survival (Time to Failure) Analysis

The randomized clinical trial is the “definitive tool” for comparing the efficacy and safety of medical and/or surgical management strategies. A quick survey of influential glaucoma clinical trials reveals that the most important analytical tool is survival/time to failure analysis, usually presented in the Kaplan-Meier formulation. Figure 1 displays the one year follow-up Kaplan-Meier curves for the Tube Versus Trabeculectomy (TVT) study. The graph gives an easily interpretable summary of the incidence of treatment failures over the course of the study’s first postoperative year. It makes clear that in the trabeculectomy group the percentage of failures increased steadily from about 1% at four months after surgery to nearly 15% at nine months. The Baerveldt group failure rate rose to just less than 4%, between about four and six months before leveling off.

- Key to this analysis is the definition of failure. This will usually be decided by clinicians in consultation with statisticians. In a surgical trial this may include re-operation and incidence of postoperative intraocular pressures considered unsafe.
- The simplest way to compare two treatments is to follow patient groups randomized to those treatments for a fixed period of time, say five years, and then test if proportions of patients failing are different in the two groups. The Kaplan-Meier analysis improves on this by examining whether the time to failure is different between treatment groups. This increases the statistical power (i.e., achieves the same statistical significance with a smaller sample size).
and analytically accounts for the virtually inevitable situation of patients having different follow-up times.

- Different follow-up times on the randomized treatment for clinical trial patients result from three circumstances: (1) a patient may experience treatment failure, (2) a patient may be lost to follow-up, or (3) a patient randomized late during study recruitment will, of course, have been followed for less time than those randomized earlier in the study. In the latter two situations, when a patient was successful at their last follow-up, but their subsequent state is unknown, we label them censored as of the time of their last visit.

- As patients fail or become censored they are removed from the calculation. This leads to a sequence of cumulative proportions over the course of post-treatment follow-up time as the total number of patients failing increases and the number of patients left at risk decreases; therefore, no single denominator applies to a time to failure analysis and no single statistic conveys all the information contained in a Kaplan-Meier curve.

- In addition to, or instead of, displaying the Kaplan-Meier curves, investigators may wish to provide cumulative proportions at specific postoperative time points, say one, three, and five years. Because these proportions, calculated at different times from randomization, do not have the same denominator they should be accompanied by their standard errors when included in the article’s text.

- Kaplan-Meier analysis is a method of summarizing time to failure data (in the same sense that an average summarizes a set of measurements), and does not itself generate a p-value. The most common significance test presented in support of a difference between Kaplan-Meier curves is the log-rank test.

- Pitfalls: (1) Kaplan-Meier analysis does not account for selection bias in who returns for follow-up or in who receives a reoperation. For example, one concern in the TVT study was whether investigators would be more likely to reoperate on trabeculectomy patients than Baerveldt patients with similar postoperative pressures. In fact examination of pressures prior to reoperation in both groups did not find this to be the case. (2) During later follow-up times the number of patients left at risk may be small and a single failure could cause a dramatic increase in the failure rate. The sample sizes of patients remaining at risk at appropriate time intervals should be supplied on the Kaplan-Meier graph or in the figure legend.

**Accounting for known Prognostic Factors and/or Studying Risk Factors: Cox Proportional Hazards Regression**

Even when comparing groups that have been randomized to treatment, investigators may wish to know if a treatment effect becomes stronger or weaker after accounting for known prognostic factors or to study the importance of putative risk factors adjusting for other important variables in a multivariate sense. Kaplan-Meier analysis allows stratification for prognostic factors, but is a univariate technique and is poorly suited to answering these questions. Cox proportional hazards regression is a flexible and popular statistical method for just this purpose.11

- The Cox model is a type of multiple regression analysis appropriate for survival data. This method allows the calculation of the relative risk of failure in a treatment group compared to a control, a group with a given risk (or prognostic) factor compared to one without, or the increase in risk of failure associated with a higher level of a continuous variable, such as an additional decade of age.

- Rather than explicitly estimating failure risks of different patient groups over time, as in Kaplan-Meier analysis, Cox analysis assumes that even if the risk of failure for two individuals receiving different treatments increases or decreases irregularly with time, their relative risk remains constant. In other words, their hazard rates are proportional.

- Specifically, the analysis gives the investigator coefficients, standard errors, and p-values assessing the importance of each variable included in the Cox model. These coefficients and standard errors can be easily transformed to risk ratios and confidence intervals.

- A risk ratio of 1.0 comparing an experimental treatment’s efficacy to a standard treatment means there is no difference between the two treatments. Risk ratios larger or smaller than 1.0, suggest an effect or difference. Thus, the OHTS treatment group risk ratio of 0.40 means that the medication group is less than ½ as likely to progress to glaucoma as the observation group.7 On the other hand in the OHTS unadjusted risk factor analysis, patients with heart disease had a 2.1 times higher risk of progression to glaucoma than those without.12 Note that in this case, the reduced risk of treatment is proportionally larger than the increased risk from having heart disease (because 1/0.4 = 2.5, which is greater than 2.1). Of course the evaluation of the treatment effect was also the reason for this prospective study, while the heart disease effect was only one of many risk factors evaluated after the fact and became non-significant in multivariate analysis.

- The interpretation of risk ratios for dichotomous variables such as treatment type and heart disease is straightforward. When estimating risks with Cox (and logistic) regression risk ratios for continuous variables, such as IOP, it becomes important to understand that the risk ratio is expressed for a particular increment of that variable. For example, the OHTS12 found that the risk of progression to glaucoma was 1.1 times higher for an individual with a baseline IOP 1 mmHg higher than another; this is expressed as a risk ratio of 1.1. It is a misconception to think that if a person has an IOP 2 or 3 mm Hg higher than another, their risk has
doubled or tripled. Because Cox regression is an exponential model, the change in risk ratio is calculated by use of exponents. If 1.1 is the risk ratio per mmHg of higher IOP, then the risk ratio for 2 mm Hg higher IOP is 1.1² (1.21). For 10 mm Hg higher IOP at baseline, the risk ratio is 1.1⁷ (=2.6), meaning that the person with 10 mm Hg higher IOP than another at baseline has a 2.6 times higher risk of progressing to glaucoma.

• Pitfalls: (1) While Cox regression analysis gives useful information about relative risks, to be properly interpreted, this must be placed in the context of actual incidence rates. For example, the OHTS risk ratio of 0.4 could be consistent with 20%/50%, 4%/10%, or 0.08%/0.2%. In fact, the second ratio is close to the actual five year OHTS incidence rates. (2) The Cox proportional hazards assumption may not fit a particular set of data. This assumption can be checked graphically or explicitly, with a significance test. Problems of this sort can be addressed with a stratified model. (3) With data sets having small sample sizes, especially when including few failures, a Cox model may not converge—that is, the computer program may be unable to correctly calculate model parameters. An indication of this is a very large estimated standard error for a variable.

**Errors Associated with Planning a Study versus Errors Analyzing the Data**

Statistical errors are divided into two types: calling a chance finding real (alpha error) or failing to find a real effect due to insufficient sample size (beta error).

• Beta error, which is 1–the statistical power of a study, is a tool for study planning. Beta is set by the investigator, usually at a level of 20%, 10%, or 5%. The choice of beta, together with an estimate of likely treatment efficacy, allows calculation of the necessary sample size, which increases as beta decreases. ¹³

• The choice of beta (i.e. the study’s power) represents a compromise between the investigator’s comfort with a particular probability of failing to find a real treatment effect and the resources they can bring to bear to collect data (i.e., ability to recruit patients). In any case, sample size calculations are almost always approximate because details of treatment efficacy are necessarily estimates (or there would be no reason to do the study) in a study’s planning stages.

• Once a study has been completed, however, the power with which it was designed is immaterial to the analysis. Even a study planned with high power (beta of 5% or 1%) still has a chance (1/20 or 1/100) of not finding an existing clinically meaningful treatment effect. Conversely a study conducted with lower power may be “lucky” and uncover an actual clinically meaningful treatment effect with a small sample size. The statement in the methods section of a clinical trial report specifying a high power is not a guarantee to the reader that the study’s conclusions are correct, but rather an affirmation that the study was well designed.

• Conducting a study with low power is problematic for two reasons. First, there is the likely potential for waste of time and resources. Second, negative studies (those with a non-significant result) designed with low power are unlikely to be published. Readers may see one small study supporting a treatment effect and not realize that there were also nineteen small studies conducted with negative results that never made it into print.¹⁴

• By convention, the upper limit of alpha error is almost always set at 0.05. In other words, to be considered statistically significant, a p-value must be less than 0.05. The p-value quantifies the probability that a treatment effect of the size observed could be due to chance. While p-values >0.05 are not considered significant, p-values less than 0.1 are generally thought to be small enough that they ought not be simply ignored and may be labeled “borderline significant”. P-values <0.01 are usually considered “highly significant” and those < 0.001 “very highly significant”. Actual p-values should be reported, not just statements of whether they are more or less than 0.05.

• A non-significant p-value could indicate that there is no treatment effect (an important result) or that the study’s sample size was too small to demonstrate a treatment effect (an ambiguous result). Similarly, a significant p-value could mean that there is a clinically important treatment effect or that the study had a large enough sample size to make significant a small or modest treatment effect. Therefore, it is not sufficient to report only p-values, in addition the likely size of a treatment effect should be explicitly reported in the form of a confidence interval.¹⁵

**Interpretation of Confidence Intervals**

Neither a high power nor a very significant (or very non-significant) p-value is a substitute for a confidence interval calculation.¹¹ An investigator should feel obliged to present the uncertainty of their result in the form of a 95% confidence interval.¹⁵ A confidence interval may be interpreted as the likely (95%) range in which a treatment effect actually falls. Table 1 displays the results of analyses comparing failure rates of three imaginary treatments to an imaginary standard treatment as well as the OHTS comparison of medication to observation groups.

<table>
<thead>
<tr>
<th>Imaginary experimental treatment</th>
<th>P-value</th>
<th>Risk ratio comparing experimental to standard treatment</th>
<th>95% confidence interval on risk ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.92</td>
<td>1.1</td>
<td>0.34-3.34</td>
</tr>
<tr>
<td>2</td>
<td>0.046</td>
<td>7.5</td>
<td>1.04-52.5</td>
</tr>
<tr>
<td>3</td>
<td>0.92</td>
<td>1.0</td>
<td>0.98-1.02</td>
</tr>
<tr>
<td>OHTS</td>
<td>&lt;0.001</td>
<td>0.4</td>
<td>0.27-0.59</td>
</tr>
</tbody>
</table>
• The comparison of experimental treatment 1 to the standard treatment provides little useful information. The risk ratio is close to one, suggesting practically no effect, but the confidence interval cannot rule out that a failure is 2.9 (1/0.34) times less likely or 3.3 times more likely in the experimental treatment group than in the standard treatment group.

• The second experimental treatment is associated with a significantly higher risk of failure than the standard treatment. The risk ratio is very large, 7.5, but the confidence interval calls into question whether the additional risk is small (4% higher in the experimental group) or extraordinarily large (53 times higher in the experimental group).

• The third comparison, though not statistically significant, gives more useful information than either of the previous two: there is no difference between the experimental and standard treatments to within ± 2 percent.

• Finally, the confidence interval for the OHTS documents that the 4/10 risk ratio comparing medication group risk to observation group risk is a good estimate and the true risk ratio is likely between 0.3 and 0.6.

• Risk ratio confidence intervals (unlike confidence intervals around averages) are symmetric on a logarithmic scale and for sensible interpretation should be displayed that way. Figure 2 presents two mathematically correct representations; however, the left hand graph with its linear scale emphasizes the portion of the confidence interval associated with increased risk, even though in fact the possible beneficial effect is almost identical.

Randomize Patients or Select a Study Eye

One of the assumptions most simple statistical methods use to arrive at standard errors and p-values is the independence of observations included in an analysis. The inclusion of two eyes of the same patient in an analysis is a clear violation of this assumption; however, the last three decades have seen development of statistical methods for analyzing correlated data and easy access to them in computer programs such as SPSS and SAS.\textsuperscript{17, 18} Notwithstanding, for ethical reasons it may still be problematic to randomize both eyes of a patient. The ability to generalize the results of a clinical trial rests on the inclusion of all eligible patients; that is, selection bias in who enters the trial must be minimized. An investigator, recognizing that a patient’s two eyes are likely to respond similarly to treatment, may well be reluctant to randomize the second eye if the first eye has already failed a study surgery. This could lead to selection bias. Of course, this is not a problem if a patient’s two eyes are randomized simultaneously.

Follow-up: Honest Accounting of Missing Data in Research Reports

• One analytical method known to be invalid is comparison of proportions at “last follow-up”. This method is prone to important selection bias related errors.\textsuperscript{19}

• Last observation carried forward analyses can, in some settings, provide insight when participants are withdrawing from a study for systematic reasons. An example is the European Glaucoma Progression study in which patients withdrew from follow-up as their intraocular pressures rose.\textsuperscript{8} However, over-interpretation of these analyses are dangerous and they are best considered as ancillary.

• Regular study follow-up visits (e.g. annual or semi-annual) are bracketed by contiguous visit windows around the ideal visit date. For example, a one-year follow-up visit might in practice include follow-up times ranging from 9 to 15 months postsurgery. Follow-up data, such as postoperative intraocular pressures and number of medications used is most honestly reported and most easily understood in the context of study visits with sample sizes indicating the number of participants lost to follow-up at each time interval.

• Studies aiming to demonstrate pre- minus post-operative differences (e.g. postoperative reductions in intraocular pressure) should report those mean differences rather than just initial and follow-up averages. The reason is that all participants are likely to contribute preoperative measurements but, almost certainly, not all of them will be measured at each follow-up visit. Those patients contributing follow-up data may well have different initial measurements from those missing follow-up. Therefore, the
Bonferroni and the Multiple Comparison Problem: Reasons not to Adjust

There are two data analysis situations referred to as “multiple comparison” problems. In the first, a test such as an analysis of variance (ANOVA) reveals a difference with respect to some measurement (e.g. intraocular pressure) among several groups of participants. In this situation, a multiple comparison method, such as the least significant difference test or multiple pairwise Bonferroni adjusted tests must be employed to determine which groups are different from each other. Approaches to dealing with this problem are not controversial. However, a very different second situation in which “multiple comparisons” are made is highly controversial. This controversy surrounds statements such as “Because each significance test is conducted with an alpha error of 0.05, 5% of the scientific literature is wrong”. Similarly, when many statistical significance tests are made within a single study it may be feared that 5% of them will be statistically significant by chance. A common prescription for this problem is to “adjust” the p-value required for significance to less than 0.05 to an extent determined by the number of significance tests performed. Two popular methods are the Bonferroni and the slightly less conservative Sidak procedures. This adjustment is usually a bad idea.

- While p-value adjusting firmly controls the rate of false positive significance tests (alpha error) it increases the number of false negative tests (beta error) – substituting one type of undesirable error for another.
- These methods address the problem of a series of independent tests (as if hypotheses in a study were so many coin tosses), but frequently variables in a study will be correlated with each other. For example, a number of GDX parameters may all be sensitive to the same underlying glaucomatous structural defects and would be incorrectly characterized as independent.
- To what family of statistical tests should the adjustment be applied? All those presented in the same table? All those presented in the same manuscript? All those studied in the same clinical trial? All those in an investigator’s career? Whatever level one chooses for adjustment favors unfortunate strategies: splitting of results over different manuscripts, or worse, not testing hypotheses at all, or worse still, conducting statistical hypotheses tests, but failing to mention the number of non-significant results that were ignored and not reported in a manuscript.

- Suppose one conducts four independent hypothesis tests and Bonferroni adjusts the statistical significance level to 0.0125. If all four individual tests had p-values of 0.03 none would reach significance, yet the chance of this outcome (four p-values < 0.05) occurring randomly is < 0.0001. Probably some true differences are being ignored (beta error).
- Science does not proceed as a string of unconfirmed answers to questions. Results are either explicitly replicated by other researchers or tested by inference to other related studies. This procedure, while not perfectly efficient, weeds out aberrant results.

A better strategy is for researchers to indicate clearly those prospective hypotheses a study was designed to evaluate as opposed to ancillary ones considered after the data were collected, and that readers (and reviewers) become “informed consumers” of p-values. A reader should place more weight on the results of the main hypotheses the study was designed to evaluate, and informally take into account the number of tests conducted when evaluating discussions and conclusions. They should view with suspicion a single significant counter-intuitive result among many tested. Still, unplanned ancillary tests – sometimes called “hypothesis generating” – are an important mechanism for identifying future research directions and should be encouraged as long as they are presented with suitable caveats.

REFERENCES


17. SPSS, Inc. 2007. Chicago, IL USA.

18. SAS Institute, Inc. Cary, NC. USA.

