Ophthalmic statistics note 2: absence of evidence is not evidence of absence

SCENARIO 1
Patients undergoing vitrectomy surgery for idiopathic full-thickness macular holes used to be routinely advised to follow a strict regime of posturing face down for a variable period (up to 2 weeks) after surgery.1–3 There was a scientific rationale for this—the tractional forces of gravity would force gases against the macula allowing it to heal more readily. Patients who postured were therefore believed to be less at risk of their macular hole reopening and of the need for repeat surgery to repair the hole. Medicine has clearly changed very significantly over time with a far greater emphasis on patient based outcomes and upon the need for an evidence base to justify practice.4–5 A senior colleague tells me that he has run a large randomised controlled clinical trial on patients who have had vitrectomies for macular holes. He states that the trial shows there is no difference in failure rates between patients who spent a week posturing face down after surgery and those who did not. He considers that this trial means that it is now unethical to ask patients to posture—particularly because several patients who did posture fed back to him how uncomfortable they found posturing. I ask him for a little more information about the trial and learn that it was a randomised controlled clinical trial with larger numbers of patients than typically found in ophthalmic surgical studies of 200 patients in each arm. Of those who spent a week posturing face down, one required repeat surgery. Of those who did not, two required repeat surgery. There is a published p value from a Fisher’s exact test that was used to compare failure rates in the two groups of 0.999 and what seems to me to be an entirely cogent argument that this demonstrates no need for posturing (see online supplementary appendix 1, table 1 for results of analysis). I have a persistent doubt however that something is not quite right with this argument and the issue leaves me pondering somewhat. I decide to go back to grass roots and search the internet for a definition of a p value.

The p value is the probability of obtaining the observed data or data that were more extreme due to chance if the null hypothesis were true.

I am somewhat perplexed by the term null hypothesis and again resort to the internet.

The null hypothesis is the situation you believe exists (in this scenario that the effect of interest is zero) and you perform a significance test to see whether there is sufficient evidence for you to reject the null hypothesis.

My interpretation of this in this scenario is that the null hypothesis is that the risk of failure with posturing following surgery is the same as the risk of failure with no posturing. Continuing in this vein, if there truly is no difference between the risks in the two groups the probability of observing the difference that I observed in this trial or something more extreme (two failures in the non-posturing group vs one failure in the posturing group) by chance alone is 0.999. I recall that p values must lie between 0 and 1, with a value of 0 meaning impossible and a value of 1 meaning absolute certainty. Here, a p value of 0.999 indicates that there is a very high chance that I would see a difference in proportions of 2/200 versus 1/200 due to chance alone and thus I have no evidence to reject the null hypothesis.

What does this mean? I have no evidence of a difference in failure rates and thus no evidence to support the use of posturing. Can I now simply advocate that it is safe for all patients not to posture? Patients have reported that they do not enjoy posturing, but the prospect of repeat surgery after an initial failure is also very daunting.

DISCUSSION
This scenario is given to illustrate challenges faced when interpreting statistical non-significance. Altman and Bland discuss this issue in a paper entitled ‘Absence of evidence is not evidence of absence’.6 Altman and Bland advocate that when presented with the statement ‘there is no evidence that’ consideration must be given as to whether absence of evidence means that there is no information at all. They suggest estimating the effect with a Confidence interval (CI) rather than simply looking at p values. Figure 1 illustrates a simple flow diagram approach based on this. Table 1 illustrates the application of the flow diagram approach to Scenario 1. In the scenario given, the odds of failure in posturing patients were 1/199, while those in the non-posturing patients were 2/198. (Odds are commonly seen in this context rather than risks, but for rare events, the OR and relative risk are approximately equal). Clearly the odds are slightly higher in the non-posturing group. The OR is 2 with a 95% CI of 0.18 to 22.3—see online supplementary appendix 1, table 2 for the computation of this. The odds of failure are estimated to be twice as common in the non-posturing patients as in the posturing patients but the CI

<table>
<thead>
<tr>
<th>Table 1 Application of flow diagram approach to Scenario 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Research question</td>
</tr>
<tr>
<td>2. Null hypothesis</td>
</tr>
<tr>
<td>3. Result</td>
</tr>
<tr>
<td>4. Interpretation</td>
</tr>
</tbody>
</table>

Figure 1 Flow diagram when presented with non-significance.
indicates considerable uncertainty in this estimate. The data are indeed consistent with there being no difference between the two trial arms in that the CI includes an OR of 1 (no difference) but the data are also consistent with the odds of failure in the posturing arm being as much as 22 times the odds of failure in the non-posturing arm (the upper limit of the CI) or indeed as little as a fifth as high. I feel much less confident now in stating that there is no difference between the treatment arms and am not sure that I entirely agree with my colleague about it being unethical to ask patients to posture when there is so much uncertainty in my estimate.

By computing a CI uncertainty is revealed which wasn’t apparent when simply looking at a p value. So should patients be posturing or not? The answer is currently unclear. What hopefully is clear is that absence of evidence is not evidence of absence and to assume that it is the case is unwise.

Most randomised trials wish to determine whether a treatment is superior to the current standard treatment. However non-inferiority and equivalence trials are becoming more common in the medical literature. A non-inferiority trial seeks to determine whether a new treatment is not worse than the standard treatment by more than an acceptable amount (known as the non-inferiority margin). An equivalence trial seeks to determine whether a new treatment is therapeutically similar to a standard treatment, that is, whether a new treatment differs from the standard treatment by no more than the non-inferiority margin. It is important to note that the term equivalence has in the past been used in error to report negative results of superiority studies—such trials often lacked statistical power to rule out important differences. The trial conducted by my colleague has not demonstrated equivalence as most clinicians and patients would consider an OR of 22.3 (which lies within the CI) as an unacceptable difference, although defining the non-inferiority margin can present a real challenge to researchers.

**SCENARIO 2**

The issue is of particular relevance when considering adverse events. These may be rare yet catastrophic for the individuals affected and their families. For treatments that are in widespread use even small differences in risk can equate to sizeable numbers of people and very large studies are needed to demonstrate differences. The recent controversy regarding the use of bevacizumab (Avastin) or ranibizumab (Lucentis) for the treatment of age-related macular degeneration (AMD), the leading cause of certifiable sight loss in the UK, very much centres around the absence of evidence issue.10 Ranibizumab was licensed for ocular use but costs substantially more than bevacizumab which does not have a marketing authorisation in this indication. Prior to licensing for AMD treatment, many people chose to have treatment with bevacizumab since without any treatment they faced rapid blindness and they preferred to accept the possibility of increased side effects with the unlicensed product. A large body of evidence built up as a result of off license use, which suggested little evidence of harm however this evidence was mostly from case series rather than Level 1 evidence. The ABC study demonstrated that bevacizumab was better than standard National Health Service (NHS) care (prior to licensing of ranibizumab) and that it appeared to offer similar benefits to ranibizumab while not appearing to increase harms.11 The study was not designed to have adequate power to examine safety concerns and so failure to detect a difference should not equate to evidence of safety. The harms under consideration were not trivial and included arteriothrombotic events and heart failure. Two large studies, inhibition of VEGF in age-related choroidal neovascularisation (IVAN) and comparison of age-related macular degeneration treatments trial (CATT), have recently been reported, both of which suggest that the drugs are indeed very similar with respect to harms and safety, and calls to license bevacizumab for use in AMD in the NHS have been made.12-14 IVAN and CATT were conducted in different parts of the world, yet the methodology was sufficiently similar to enable results from the two studies to be validly combined using a technique called meta (Greek for after) analysis. Recent pooling of results from the two studies has suggested that a higher proportion of patients who receive bevacizumab experience one or more serious adverse events, although numbers are small and so the jury is still out. Table 2 illustrates the application of the flow diagram approach to Scenario 2.
REFERENCES


