Do not throw the baby out with the bath water: a guide for using non-significant results in practice

Mouaffaa Tello,[1] Feras Zaiem,[1] Mary Catherine Tolcher,[1,2] Mohammad Hassan Murad[1]

10.1136/ebmed-2016-110510

[1]Evidence-based Practice Center, Mayo Clinic, Rochester, Minnesota, USA
[2]Department of Obstetrics and Gynecology, Division of Maternal-Fetal Medicine, Baylor College of Medicine, Houston, Texas, USA

Correspondence to:
Dr Mohammad Hassan Murad,
Evidence-based Practice Center,
Mayo Clinic, 200 First Street SW,
Rochester, MN 55905, USA;
murad.mohammad@mayo.edu

Précis Statistically non-significant results can sometimes be clinically useful and help in decision-making.

Abstract

Acting on results that are not statistically significant is challenging for clinicians. Such results are often interpreted as evidence of lack of association or as useless evidence. We provide a framework for interpreting and applying non-significant results at the point of care using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach.

It is well known that p values are misused, misunderstood and miscommunicated.[1] Much has been written about misleading conclusions based on p values and data dredging; and how this contributes to publication bias and misleading conclusions. However, there is minimal guidance for clinicians interested in practicing evidence-based medicine on how to actually implement results that are not statistically significant in patient care. Many scientific publications have cautioned about the limitations of p values and provided the correct definition and interpretation of them, but a practical guide is needed. To merely tell practitioners that p value is defined as the probability of observing events as extreme or more extreme than the observed data, given that the null hypothesis is true (which is the correct definition), is not very helpful.

We provide a framework for interpreting and applying non-significant results at the point of care with an example. This framework is derived from the fields of statistics, evidence-based medicine and patient-centred shared decision-making and is implemented using the GRADE approach.[2]

Example

A study of women with a history of prior caesarean compared a trial of labour versus a repeat caesarean delivery.[3] The study showed that neonatal death did not significantly differ between the two approaches (OR 1.82, 95% CI 0.73 to 4.57; p=0.19). One can simply conclude that both modes of delivery are equivalent and offer them to women as equal options without further scrutiny. Others may consider this evidence to be useless and cannot be used in practice. Both approaches are potentially misguided.

Using non-significant results in practice

Estimate the absolute effect

For making a decision, the absolute effect is needed. The p values and relative effects cannot be directly communicated to a patient or a policymaker.[4] Calculating the absolute effect can be carried out even if it was not reported in the published manuscript. This can be carried out simply by subtracting the control event rate from the intervention event rate (trial of labour 13/15 338 (0.085%), elective caesarean delivery 7/15 014 (0.047%), the difference is four more neonatal deaths per 10 000 deliveries).

Plausible limits for this effect can be calculated using various methods,[5 6] some of which do not require a statistical software package (simple calculation using the formula in box 1 or using Grading of Recommendations Assessment, Development and Evaluation (GRADE) online applications). An OR and a control event rate are the only two values required for such calculations. In this example, the risk of neonatal death associated with a trial of labour has a range of plausible values from 2 less to 20 more deaths (per 10 000 deliveries). Per 1000 deliveries (the usual denominator in GRADE evidence profiles used for decision-making and guideline development), the increase in neonatal deaths would range from 0 to 2 more. This absolute effect remains statistically non-significant; however, it may allow a decision to be made.

Rate certainty in the evidence

The GRADE approach uses eight different domains to rate the certainty in evidence (also called quality of evidence); however, since this discussion is about statistically non-significant results, we focus on the domain of imprecision.

GRADE advises decision-makers to ignore p values and instead determine whether their decision would change across the CI of the absolute effect.[7]

In this scenario, if we consider a neonatal death rate of 2 per 1000 to be trivial, then our decision would be the same whether the true effect was 0/1000 or 2/1000. In this case, this evidence would be precise and warrants sufficient certainty to make a decision. If we consider a neonatal death rate of 2/1000 to be unacceptable, then this evidence is imprecise and we have very low certainty about making a decision. The second layer of

Box 1 Calculating an absolute effect* from relative association measures

1. From a relative risk:

$$(1 - \text{relative risk}) \times \text{CER}$$

2. From an OR:

$$\text{CER} - \frac{\text{OR} \times \text{CER}}{1 - \text{CER} + \text{OR} \times \text{CER}}$$

*Also called risk difference or absolute risk reduction.
CER is the control event rate.

decision-making involves a fully contextualised approach that incorporates other outcomes. For example, some women may consider the increase in neonatal death to be acceptable if they knew that a trial of labour is associated with lower risk of maternal thromboembolic complications.

Use shared decision-making principles and tools

The second principle of evidence-based medicine dictates that evidence alone is insufficient for decision-making. One should also consider patients' values and preferences. In the example discussed so far, clinicians should ask women how they value each outcome. Women who fear (are more averse to) any fetal loss even if the risk was very low may opt for a caesarean delivery. Women who place higher values on preventing maternal perioperative complications, prefer a natural birth experience and short postdelivery recovery period may opt for a trial of labour considering that the risk of fetal loss is very low.

Clinicians presenting these options should convey the uncertainty in evidence to patients. Shared decision-making processes and tools, such as decision aids, are paramount to facilitate the discussion and convey probabilities using natural frequencies (which are more understandable) and depictions. Such tools have been associated with lower decisional conflict, lower chances of regret and increased likelihood that patients will make decisions consistent with their own values.[8 9]

Discussion and conclusion

Text mining studies have demonstrated that the use of p values has substantially increased from 1990 to 2014.[1] Therefore, this issue will continue to be a challenge for evidence-based practitioners. Statistically non-significant results can be clinically useful and may help with making clinical decisions. Avoiding harm was presented here as an example. Evidence-based healthcare practice hinges on using the best available evidence; if that evidence was statistically non-significant, we have a compelling rationale to use it.

Bayesian approaches may be a better alternative that hypothesis testing for inference because they involve the calculation of the probability of parameters given the data (as opposed to the frequentist approach; which computes the probability of the data given the parameters). Therefore, Bayesian approaches can directly provide the posterior odds of the null hypothesis against the alternative hypotheses (eg, 4:1 being true).[10] These odds may be more intuitive to decision-makers than the p value which is commonly misinterpreted as 'the probability of the null hypothesis being true'.

In this methodological proposal, we provide a framework for interpreting and applying non-significant results that incorporates statistics, evidence-based medicine and shared decision-making perspectives.

Contributors MHM and MCT conceived the idea of this guide and wrote the initial draft. FZ and MT revised the manuscript and contributed to the discussion.

Competing interests None declared.

Provenance and peer review Not commissioned; internally peer reviewed.

References

1. **Chavalarias D**, Wallach JD, Li AH, *et al*. Evolution of reporting p values in the biomedical literature, 1990–2015. *JAMA* 2016;**315**:1141–8.
2. **Guyatt GH**, Oxman AD, Kunz R, *et al*. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008;**336**:995–8.
3. **Landon MB**, Hauth JC, Leveno KJ, *et al*. Maternal and perinatal outcomes associated with a trial of labor after prior cesarean delivery. *N Engl J Med* 2004;**351**:2581–9.
4. **Murad MH**, Montori VM, Ioannidis JP, *et al*. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. *JAMA* 2014;**312**:171–9.
5. **Guyatt GH**, Eikelboom JW, Gould MK, *et al*. Approach to outcome measurement in the prevention of thrombosis in surgical and medical patients: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;**141**(Suppl 2):e185S–94S.
6. **Murad MH**, Montori VM, Walter SD, *et al*. Estimating risk difference from relative association measures in meta-analysis can infrequently pose interpretational challenges. *J Clin Epidemiol* 2009;**62**:865–7.
7. **Guyatt GH**, Oxman AD, Kunz R, *et al*. GRADE guidelines 6. Rating the quality of evidence–imprecision. *J Clin Epidemiol* 2011;**64**:1283–93.
8. **Knops AM**, Legemate DA, Goossens A, *et al*. Decision aids for patients facing a surgical treatment decision: a systematic review and meta-analysis. *Ann Surg* 2013;**257**:860–6.
9. **Stacey D**, Bennett CL, Barry MJ, *et al*. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev* 2011(10):CD001431.
10. **Lee JJ**. Demystify statistical significance–time to move on from the p value to bayesian analysis. *J Natl Cancer Inst* 2011;**103**:2–3.