The fragile p-value

Abdus Samad Ansari explores the limitation of the p-value and the application of the fragility index in clinical trials.

Clinical trials and tribulations?

The restoration of vision or more purely the gift of sight is an aspect of care that provides an ophthalmologist with the ability to change the life of a patient. This fundamental pillar is what drives the research we do, the exams we sit and the hours we practise surgical technique. The introduction of novel revolutionary therapies has dramatically changed the landscape of this field of medicine. Yet we strive for more: better outcomes, safer procedures, reduced costs: all in the name of the patient.

Information retrieved from the highest quality evidence, most often findings from randomised controlled trials (RCTS), is used to inform healthcare decisions at both the population and individual level. Clinicians and researchers alike spend considerable efforts conducting trials with the intent of establishing treatment effectiveness and altogether improving patient outcomes. From development of research questions to decisions regarding 'significant' treatment targets, researchers control evidence generation and interpretation in its entirety. The criteria which guide our decisions to pursue research objectives are weighted heavily in: 1) clinical equipoise regarding the intervention, and 2) the ethical justification and established benefit findings from such research may exemplify [1].

Clinical equipoise and the ethical justification

Informed decision-making concerning the choice of intervention requires us to rely on evidence-based guidelines, most often supported by an abundance of literature comprising meta-analyses and randomised controlled trials. Whenever a situation arises, where an evidence-based choice of intervention is not possible, the assumption of equipoise surfaces. Clinical equipoise is the guiding principle, which rationalises the completion of randomised trials within human populations; it requires the investigators to be in a state of genuine uncertainty when contemplating an intervention's effectiveness for treating disease [1,2].

Uncertainty regarding an intervention's effectiveness is compulsory to justify the random allocation of treatment and placebo to sick populations. There is no question concerning the appropriateness of conducting RCTs to establish the efficacy of interventions for dry age-related macular degeneration (AMD) – for currently this disease has no recognised therapies which halt or reverse progression. Thus, any pursuit of research regarding this question maintains the principle of clinical equipoise. Conversely, do you believe there is any question of the therapeutic merit surgical intervention provides in the treatment of cataracts? Two instances exist which permit the use of placebo; the first being a case when withholding therapy poses negligible risk, and the second being a case when no effective treatment currently exists.

Evidence based interpretation

Clinical trials which determine our clinical practice must adhere to a number of systems to ensure they merit publication in a high impact journal. They must be well designed and have achieved the appropriate sample size and power calculations, thus ensuring we can accurately interpret results [3]. Since medical school, clinicians are taught about the p-value. Its use in denoting statistical significance has long been the benchmark by which results are deemed to be valid. However, these can often be inappropriately applied or misinterpreted. The development of the Consolidated Standards of Reporting Trials (CONSORT) now encourages not only the use of the p-values but also confidence intervals to determine true magnitude of effect and precision of results [4]. However, despite this information one cannot easily discern the reproducibility of results if the trial was to happen again.

> **"The Fragility Index is a tool that allows us to gauge the robustness or 'fragility' of the results of a clinical trial – the number of events required to convert a trial from being statistically significant to not statistically significant"**

The Fragility Index

In 2014 Walsh et al. looked to gauge how fragile this p-value truly was, with the subsequent creation of the Fragility Index (FI) [5]. They questioned the concept of statistical significance and its implications that a p-value of less than 0.05 implies that an observed result is unlikely to have occurred by chance alone. It was believed that many readers often place a similar degree of belief in results of a trial based on the p-value, irrespective of other factors such as size of the trial or the number of events that take place [5]. They looked to evaluate methods to better communicate the limits of the p-value by creating a new metric that would be able to demonstrate how easily significance could be lost. In order to do so, they identified RCTs with statistically significant results in high-impact journals (*The Lancet, New England Journal of Medicine, the British Medical Journal etc.*) that had at least one dichotomous outcome (e.g. myocardial infarct, graft failure, best-corrected visual acuity >6/18, etc.) or time to event outcome in the study's abstract [5]. By changing the status in the group with the smallest number of events they changed the outcome from no event to a positive finding and recalculated the analysis until they exceeded a p-value of 0.05 [5]. This variation was labelled the FI, with a smaller number proposing a more fragile result. From the 399 eligible trials the median sample size was noted to be 682 patients with 53% reporting a p-value <0.01. They calculated the median fragility to be nine (range 0-109) with a quarter of trials having a FI of three or less. In over half of the trials this FI metric was less than the number of patients that were lost to follow-up. They were able to demonstrate the statistical significance of results was indeed reliant on a small number of events and the implementation of this new metric could be a simple method to assist in the detection of less robust results. In effect, FI is a tool that allows us to gauge the robustness or 'fragility' of the results of a clinical trial – the number of events required to convert a trial from being statistically significant to not statistically significant.

For example, say we have a randomised control trial set up to minimise bias and aimed to evaluate a medication for the prevention of wet AMD. This trial consisted of 300 patients who were randomised appropriately to receive medication A or placebo. If the outcomes of the trial were compared between groups and produced a statistically significant p-value – it might appear that a true effect might exist. However, this could be influenced by a small change in the number of events – say one or two, thus losing significance and changing the way we interpret results all together. Practically this minimal change in events could be due to missing data, patients lost to follow-up / missed appointments or even simply reflect group imbalances. In fact it has long been suggested results from RCTs may be fragile [6,7] and since the creation of this powerful and intuitive concept many have looked to evaluate the results from major trials within their own specialties including oncology [8], peadiatrics [3], cardiology [9], surgery [10], anaesthetics [11] and even ophthalmology [12].

One interesting real-life example mentioned by Walsh et al. was the Leicester Intravenous Magnesium Intervention Trial (LIMIT-2) [13] published in *The Lancet*. This clinical trial was carried out to look at the effect of IV magnesium on 28-day survival in patients with acute heart attacks. The RCT had 2316 patients and it was able to show a 24% relative risk reduction in mortality with an associated p-value of 0.04. A few years later, a trial with over 50,000 patients concluded that in fact there was no merits to its use [14]. Interestingly, the FI of the LIMIT-2 trial was calculated to be one, posing the question, would these clinicians have interpreted the results in a similar manner if they knew one event could have changed the result?

> **"Would these clinicians have interpreted the results in a similar manner if they knew one event could have changed the result?"**

The fragility of trials in ophthalmology

To date, only one systematic review assessing the fragility of trials within ophthalmology has been completed [12]. It hypothesised that many of the trials within our specialty are often limited by trial size and would demonstrate a similar fragility to other specialties. Aiming to explore the robustness of these results, the review attempted to identify factors which could potentially influence FI in ophthalmology. A total of 156 trials were identified that met their inclusion criteria constructed around study design and reporting methods. Information was collected on outcomes, p-values, number of patients in each group, event rates and number of patients lost to follow-up. Of the trials evaluated, the median sample size was 91.5 (range 4-2217) with a median number of events noted to be 28 (IQR.25-65.75) [12]. The median number of missing patient data was four. The number of missing data on patients was greater than the fragility of that trial in over 50% of the included trials. They established that on average the FI of trials in ophthalmology was two (range 0-48) implying that by simply changing two non-events to events in the treatment group the results would lose statistical significance. Additionally, they noted in the major trials published, the power calculation was only reported 71.8% of the time. This underpowering of trials, particularly when effect sizes can be small would propose many underpowered trials indicating statistical significance when disparities can merely be due to chance – confirmed with the application of the FI. Features predictive of FI in ophthalmology were noted to be the p-value, sample size and the total number of events [12]. They noted that the FI is not without its own limitations. It can only be used for dichotomous outcomes in a 1:1 study design [12]. There is potential variability and limitations in its systemic application of different units of measure across trials. Additionally, trials employing more advanced survival statistical analysis would make its application more problematic. Nonetheless, these findings must be appreciated against the practicalities of completing a clinical trial. These require substantial resource allocation, significant amounts of funding and are created within the confinement of patient recruitment [15]. Thus, the ability to complete large, multicentered, adequately powered trials is not always possible. A small FI does not necessarily mean that the estimated effect of an outcome is wrong, nor does it devalue the results of trials completed – many of which cannot be replicated.

Final thoughts

The method by which we interpret trial data should be done within the context of numerous factors. This includes the effect size seen, biological plausibility, generalisability, risk of systemic bias and even conflicts of interest [12,16]. The creation of the fragility index was not to criticise the merits of the p-value but augment its interpretation. To help us make sense of the data provided to us and to question its real-world application. It is far too easy for clinicians to simply pay lip service to the statistical methods put in front of us when trying to interpret data and, with the misguided assumption it is entirely comprised of the notion that we have now utilised the 'best external evidence' when making clinical decisions. However, we must remember evidence-based medicine is composed of three core principles: 1) the utilisation of best external evidence, 2) individual expertise and 3) patient values and preferences, the sum of which is only possible if we critically evaluate the information given to us and appreciate which results necessitate further assessment.

References

1. Cook C, Sheets C. Clinical equipoise and personal equipoise: two necessary ingredients for reducing bias in manual therapy trials. *J Man Manip Ther* 2011;**19**:55-7.
2. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;**317**:141-5.
3. Matics TJ, Khan N, Jani P, Kane JM. The Fragility Index in a Cohort of Pediatric Randomized Controlled Trials. *J Clin Med* 2017;**6**:79.
4. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010;**152**:726-32.
5. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. *J Clin Epidemiol* 2014;**67**:622-8.
6. Feinstein AR. The unit fragility index: an additional appraisal of "statistical significance" for a contrast of two proportions. *J Clin Epidemiol* 1990;**43**:201-9.
7. Walter SD. Statistical significance and fragility criteria for assessing a difference of two proportions. *J Clin Epidemiol* 1991;**44**:1373-8.
8. Del Paggio JC, Tannock IF. The fragility of phase 3 trials supporting FDA-approved anticancer medicines: a retrospective analysis. *The Lancet Oncology* 2019;**20**:1065-9.
9. Khan MS, Ochani RK, Shaikh A, et al. Fragility Index in Cardiovascular Randomized Controlled Trials. *Circulation: Cardiovascular Quality and Outcomes* 2019;**12**:e005755.
10. Evaniew N, Files C, Smith C, et al. The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey. *The Spine Journal* 2015;**15**:2188-97.
11. Bertaggia L, Baiardo Redaelli M, Lembo R, et al. The Fragility Index in peri-operative randomised trials that reported significant mortality effects in adults. *Anaesthesia* 2019;**74**:1057-60.
12. Shen C, Shamsudeen I, Farrokhyar F, Sabri K. Fragility of results in ophthalmology randomized controlled trials: a systematic review. *Ophthalmology* 2018;**125**:642-8.

13. Woods KL, Fletcher S, Roffe C, Haider Y. Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2). *The Lancet* 1992;**339**:1553-8.
14. Collins R, Peto R, Flather M, et al. ISIS-4-A randomised factorial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58.050 patient with suspected acute myocardial-infarction. *Lancet* 1995;**345**:669-85.
15. Jones CW, Platts-Mills TF. Understanding commonly encountered limitations in clinical research: an emergency medicine resident's perspective. *Ann Emerg Med* 2012;**59**:425-31. e411.
16. Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. *Methods guide for effectiveness and comparative effectiveness reviews [Internet]:* Agency for Healthcare Research and Quality (US); 2012.

TAKE HOME MESSAGE

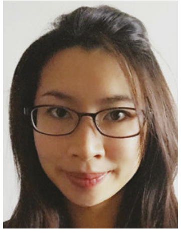
- Research objectives are weighted heavily in the clinical equipoise regarding the intervention and its ethical justification.
- Many clinicians often place a similar degree of belief in results of a trial based on the p-value, irrespective of other factors such as size of the trial or the number of events that take place.
- The implementation of the Fragility Index can be a simple method to assist in the determination of less robust results from trial data.
- Trial data should be interpreted within the context of numerous factors including the effect size seen, biologic plausibility, generalisability, risk of systemic bias and even conflicts of interest.

AUTHOR

Abdus Samad Ansari,

NIHR Academic Clinical Fellow. Ophthalmology ST4,
South London Deanery.

SECTION EDITOR



Annie SeeWah Tung,

Ophthalmology Specialty Trainee Year 7,
Wales Deanery, UK.

E: annieswtung@doctors.org.uk