EDITORIAL

There is still a place for significance testing in clinical trials

Much hand-wringing has been stimulated by the reflection that reports of clinical studies often misinterpret and misrepresent the findings of the statistical analyses. Recent proposals to address these concerns have included abandoning p-values and much of the traditional classical approach to statistical inference,[1] or dropping the concept of statistical significance while still allowing some place for p-values.[2] How should we in the clinical trials community respond to these concerns? Responses may vary from bemusement, pity for our colleagues working in the wilderness outside the relatively protected environment of clinical trials, to unease about the implications for those of us engaged in clinical trials.

We believe there is validity to the fundamental concerns that have led to these proposals about the ways in which many scientists use, interpret and report statistical tests, and that these concerns have implications for the design, conduct, analysis, interpretation, and dissemination of clinical trials. A typical clinical trial will be premised upon a statistical hypothesis testing framework, both for the determination of the sample size, and for its analysis.[3] We know that misrepresentation of findings is not something from which clinical trials are immune.[4] Misinterpretation of p-values is commonplace, and terms like "statistical significance" sometimes seem to be used with little consideration for the implications.

However, we should not be shy about asserting the unique role that clinical trials play in scientific research. A clinical trial is a much safer context within which to carry out a statistical test than most other settings. Properly designed and executed clinical trials have opportunities and safeguards that other types of research do not typically possess, such as protocolisation of study design, scientific review prior to commencement, prospective data collection, trial registration, specification of outcomes of interest including, importantly, a primary outcome, and others. For randomised trials, there is even more protection of scientific validity provided by the randomisation of the interventions being compared. It would be a mistake to allow the tail to wag the dog by being overly influenced by flawed statistical inferences that commonly occur in less carefully planned settings. Furthermore, the research question addressed by clinical trials (comparing alternative strategies) fits well with such an approach, and the corresponding decision-making settings (e.g. regulatory agencies, data and safety monitoring committees and clinical guideline bodies) are often ones within which statistical experts are available to guide interpretation.

The carefully designed clinical trial based on a traditional statistical testing framework has served as the benchmark for many decades. It enjoys broad support in both the academic and policy communities. There is no competing paradigm that has to date achieved such broad support. The proposals for abandoning p-values altogether often suggest adopting the exclusive use of Bayesian methods. For these proposals to be convincing it is essential their presumed superior attributes be demonstrated without sacrificing the clear merits of the traditional framework. Many of us have dabbled with Bayesian approaches and find them to be useful for certain aspects of clinical trial design and analysis, but still tend to default to the conventional approach notwithstanding its limitations. While attractive in principle, the reality of regularly using Bayesian approaches on important clinical trials has been substantially less appealing – hence their lack of widespread uptake.

The issues that have led to the criticisms of conventional statistical testing are of much greater concern where statistical inferences are derived from observational data. The proliferation of large, complex data sources that offer the opportunity for running a multitude of statistical analyses, often of an unplanned

and exploratory nature, leads naturally to the identification of false positive findings at a vastly greater frequency than the significance levels of the tests used would imply. Such strategies can radically undermine the probabilistic validity of the inferences. Even when the study is appropriately designed there is also a common converse misinterpretation of statistical tests whereby the investigator incorrectly infers and reports that a non-significant finding conclusively demonstrates no effect. However, it is important to recognize that an appropriately designed and powered clinical trial enables the investigators to potentially conclude there is "no meaningful effect" for the principal analysis. More generally these problems are largely due to the fact that many individuals who perform statistical analyses are not sufficiently trained in statistics. It is naïve to suggest that banning statistical testing and replacing it with greater use of confidence intervals, or Bayesian methods, or whatever, will resolve any of these widespread interpretive problems. Even the more modest proposal of dropping the concept of "statistical significance" when conducting statistical tests could make things worse. By removing the prespecified significance level, typically 5%, interpretation could become completely arbitrary. It will also not stop data-dredging, selective reporting or the numerous other ways in which data analytic strategies can result in grossly misleading conclusions.

These considerations notwithstanding, the field of clinical trials is in rapid evolution and it is entirely possible and appropriate that the statistical framework used for their evaluation must also change. However, such evolution should emerge from careful methodological research and open-minded, self-critical enquiry. We earnestly hope that *Clinical Trials* will continue to be seen as a natural academic home for exploration and debate about alternative statistical frameworks for making inferences from clinical trials.[5] The Editors welcome articles that evaluate or debate the merits of such alternative paradigms along with the conventional one within the context of clinical trials. Especially welcome are exemplar trial articles, and those which are illustrated using practical examples from clinical trials that permit a realistic evaluation of the strengths and weaknesses of the approach.

Jonathan A Cook
Dean A Fergusson
Ian Ford
Mithat Gonen
Jonathan Kimmelman
Edward L Korn
Colin B Begg

References:

1. Gill J. Comments from the new editor. *Policy Analysis* 2018;26(1):1-2
2. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-7.
3. Cook JA, Julious SA, Sones W, et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ* 2018; 363. doi: https://doi.org/10.1136/bmj.k3750.
4. Boutron I. Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Jama*. 2010;303(20):2058-64.
5. Ellenberg SS, Ellenberg JH. Proceedings of the University of Pennsylvania ninth annual conference on statistical issues in clinical trials: Where are we with adaptive clinical trial designs? *Clinical Trials*. 2017;14(5):415-6.