RESEARCH ARTICLE

REVISED Smaller clinical trials for decision making; a case study to show p-values are costly [version 2; peer review: 1 approved, 2 approved with reservations]

Previously titled: Smaller clinical trials for decision making; using p-values could be costly

Nicholas Graves[1], Adrian G. Barnett[1], Edward Burn[2], David Cook[3]

[1]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD, 4059, Australia
[2]Nuffield Department of Orthopaedics, Oxford University, Oxford, OX3 7LD, UK
[3]Princess Alexandra Hospital, Brisbane, Brisbane, QLD, 4102, Australia

v2 **First published:** 02 Aug 2018, **7**:1176
https://doi.org/10.12688/f1000research.15522.1
Latest published: 27 Sep 2018, **7**:1176
https://doi.org/10.12688/f1000research.15522.2

Abstract

Background: Clinical trials might be larger than needed because arbitrary levels of statistical confidence are sought in the results. Traditional sample size calculations ignore the marginal value of the information collected for decision making. The statistical hypothesis testing objective is misaligned with the goal of generating information necessary for decision-making. The aim of the present study was to show that for a case study clinical trial designed to test a prior hypothesis against an arbitrary threshold of confidence more participants were recruited than needed to make a good decision about adoption.
Methods: We used data from a recent RCT powered for traditional rules of statistical significance. The data were also used for an economic analysis to show the intervention led to cost-savings and improved health outcomes. Adoption represented a sensible investment for decision-makers. We examined the effect of reducing the trial's sample size on the results of the statistical hypothesis-testing analysis and the conclusions that would be drawn by decision-makers reading the economic analysis.
Results: As the sample size reduced it became more likely that the null hypothesis of no difference in the primary outcome between groups would fail to be rejected. For decision-makers reading the economic analysis, reducing the sample size had little effect on the conclusion about whether to adopt the intervention. There was always high probability the intervention reduced costs and improved health.
Conclusions: Decision makers managing health services are largely invariant to the sample size of the primary trial and the arbitrary p-value of 0.05. If the goal is to make a good decision about whether the intervention should be adopted widely, then that could have been achieved with a much smaller trial. It is plausible that hundreds of

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 2 (revision) 27 Sep 2018	? view	✓ view	? view
version 1 02 Aug 2018	? view		

1. **Stephen Senn**, Luxembourg Institute Of Health, Strassen, Luxembourg
University of Sheffield, Sheffield, UK
2. **Steven A Julious**, University of Sheffield, Sheffield, UK
3. **Daniel Benjamin Mark**, Duke University Medical Center, Durham, USA

Any reports and responses or comments on the article can be found at the end of the article.

millions of research dollars are wasted each year recruiting more participants than required for RCTs.

Keywords
decision making, RCT, sample size, waste in research

Corresponding author: Nicholas Graves (n.graves@qut.edu.au)

Author roles: Graves N: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Barnett AG**: Conceptualization, Formal Analysis, Methodology, Software, Supervision, Writing – Review & Editing; **Burn E**: Conceptualization, Data Curation, Formal Analysis, Methodology, Writing – Review & Editing; **Cook D**: Conceptualization, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Graves N, Barnett AG, Burn E and Cook D. **Smaller clinical trials for decision making; a case study to show p-values are costly [version 2; peer review: 1 approved, 2 approved with reservations]** F1000Research 2018, **7**:1176 https://doi.org/10.12688/f1000research.15522.2

First published: 02 Aug 2018, **7**:1176 https://doi.org/10.12688/f1000research.15522.1

REVISED **Amendments from Version 1**

This is an update in response to the review from Stephen Senn. The title has been changed, there is a new Figure 3, we have added a new “Counter-exmple of no treatment” sub-heading in the Methods section, and have expanded the Discussions section.

See referee reports

Introduction

Informed patients, thoughtful clinicians and rational health planners make decisions about the services and treatments provided using the best information available, and all decisions are made under conditions of uncertainty[1,2]. We examine a situation where sufficient evidence arises from a clinical trial to inform a decision about changing services before the conventional statistical stopping point for a clinical trial is reached. This paper is about the tension between the ‘precision’ and the ‘impact’ of a scientific measurement[3] and how that tension might dictate the sample size of a clinical trial.

Imagine a new treatment is compared against the best contemporary alternative in a well conducted randomised controlled trial (RCT). The design requires 800 participants in total based on a standard sample size calculation of 5% type 1 error and 80% power. The new treatment is more efficacious, prolongs life of high quality and saves more money than it costs to implement. The evidence to support these conclusions can be seen in the data after only 200 trial participants have been recruited, but primary outcomes are not yet statistically significant. Clinical equipoise, the cornerstone of ethical treatment allocation is lost, yet the conventions of hypothesis testing and arbitrary power calculation demand a further 600 participants are recruited. The information arising from the additional 600 participants is unlikely to change the actions of a rational decision maker who wishes to adopt the new treatment. Yet scarce research funds are used up meaning opportunities to fund other research are lost, and some patients have been consented and allocated to a treatment that we could not recommend, nor would we chose for ourselves or our families.

The utility of clinical trials for those managing health services and making clinical decisions is under debate and traditional paradigms are being challenged[4]. The chief claim of this paper is that an RCT designed to test a hypothesis using traditional rules of inference might have more participants than required, if the goal is to make a good decision. Waste in research arises from routine use of arbitrary levels of statistical confidence[5] and because the trial data are considered in isolation[6]. The marginal value of the information acquired for the purpose of making a good decision is not made explicit. Important information for the purpose of decision making often lies outside the clinical trial process. The plausibility of our claim is demonstrated by re-analysing a recent RCT[7].

Choosing a sample size for hypothesis testing

For the design of superiority trial, the aim is to have a high likelihood of sufficient evidence to confidently reject a null hypothesis that two treatments are equivalent when treatments differ by a specified difference. This difference is usually based on either clinical importance or a best guess of the true treatment effect. Inference based on this approach has two types of potential errors. A false-positive or type I error of rejecting the null hypothesis when there is no difference, with probability α. A false negative or type II error of not rejecting the null hypothesis when there is an effect, with probability β. The sample size of the trial is calculated to give an acceptable type I error rate and power ($1-\beta$), typically 0.05 for α and 0.8 to 0.9 for the power. The final analysis summarises the incompatibility between the data and the null hypothesis[8]. If the p-value is below the standard 5% limit the null hypothesis of no effect is rejected. A ‘statistically significant’ result is then celebrated and typically used to support a decision to make a change to health services.

Choosing a sample size for decision making

We assume the objective of decision-makers who manages health services is to improve outcomes for the populations they serve. Because this challenge will be addressed with finite resources not every service or new technology can be made available for a population. Decision-makers therefore require knowledge of the health foregone from not funding services displaced by the services that are funded[9]. The services that are provided should generate more health benefits per dollar of cost when compared to those that are not. With this criterion satisfied the opportunity cost from the services not provided is minimised. A rational decision maker will logically follow these rules: do not adopt programmes that worsen health outcomes and increase cost; adopt programmes that improve health outcomes and decrease costs; and, when they face a situation of increased cost for increased health outcomes they prioritise programmes that provide additional health benefits for the lowest extra cost[10]. They will continue choosing cost-effective services until available health budgets are exhausted. An appropriate and generic measure of health benefit is the quality adjusted life year (QALY)[11]. While this approach does not consider how health benefits are distributed among the population there is a framework for including health inequalities in the economic assessment of health care programmes[12].

In choosing a sample size for a clinical trial to evaluate a new service or technology a decision-maker will consider the uncertainty in the conclusion about how costs and health benefits change by adoption. The aim is to reduce the likelihood of making the wrong decision. They will make rational and good decisions, and they will manage uncertainty rather than demand an arbitrarily high probability of rejecting a null hypothesis. Methods are available to estimate the expected value of information and so the optimal sample size for a trial is dependent on the context specific costs and benefits of acquiring extra information[13]. Each decision is context dependent and the ‘one size fits all’ approach to sample size calculation is arbitrary and potentially wasteful. This holistic approach should be a priority for designing, monitoring and analysing clinical trials.

Methods

The TEXT ME RCT: A case study

A case study to illustrate the differing evidential requirements of the ‘hypothesis-testing’ and ‘decision-making’ approaches is provided by the RCT of the Tobacco, Exercise and Diet

Messages (TEXT ME) intervention[14]. This health services program targeted multiple influential risk factors in patients with coronary heart disease, with SMS text messages. Advice and motivation was provided to improve health behaviours and it was supplementary to usual care. The hypothesis was that the intervention would lower plasma low-density lipoprotein cholesterol by 4.5 mg/dL at 6 months for participants compared with those receiving usual care[15]. The required sample size was 704 participants for 90% power[15] and the trial recruited and randomised 710 participants[7]. The mean difference between the intervention and control group was –5 mg/dL, (95% CI –9 to 0 mg/dL). With a p-value of 0.04, the null hypothesis was rejected. Evidence for health effects were also sought on other biomedical and behavioural risk factors, quality of life, primary care use and re-hospitalisations. Clinically and statistically significant effects were also found for systolic blood pressure (mean difference –8 mmHg, p<0.001), body mass index (–1.3 kg/m^2, p<0.001) and current smoking (relative risk of 0.61, p<0.001).

The TEXT ME trial data were used to inform an economic evaluation of the potential change to costs and health benefits measured in quality adjusted life years to the community from a decision to adopt the programme[16]. The observed differences in low-density lipoprotein cholesterol, systolic blood pressure and smoking were combined with reliable external epidemiological evidence to estimate the reduction in acute coronary events, myocardial infarction and stroke and were extrapolated over the patients expected remaining life times. The costs of providing the intervention, the projected costs of the treatment of acute events and general primary care use and expected mortality were all informed by data sources external to the primary trial[16]. The findings revealed that TEXT ME was certainly going to lead to better health outcomes and cost savings. The conclusion was that a rational decision-maker should fund and implement the TEXT ME program. Once available an informed clinician would then recommend TEXT ME to coronary patients, and enough patients would sign up to create benefits for individuals and the health system. Using the TEXT ME study, we consider whether the same decision could have been made at an earlier stage with fewer participants enrolled in the primary trial.

Data analysis

We examine the effect of a reduced sample size on the results of both the hypothesis-testing analysis for differences in low-density lipoprotein cholesterol, and the economic evaluation of the intervention. From the original 710 participants, smaller samples between 100 and 700 patients in increments of 100 were considered with the resampling done with replacement. The 'p-value' and 'economic' analyses were re-run using the data provided by the randomly selected patients and this process was repeated 500 times for each sample size. The simulations and figures were created using R (version 3.1.0). The code is available on GitHub https://github.com/agbarnett/smaller.trials but we are unable to share the primary data from the TEXT ME RCT.

Counter-example of no treatment effect

To illustrate this approach with treatments that are equally effective, we used the same methods as above, but created data using the TEXT ME trial where the two groups had equivalent outcomes. We did this by randomly allocating patients to the TEXT ME intervention or usual care, and then resampling with replacement to create a new version of the study sample. We assumed there was no risk reduction for the TEXT ME group, and used the same uncertainty in risk reduction as per the previous model.

Results

The effect of reducing the sample size for hypothesis-testing objectives was to simulate studies that traditional hypothesis testing approaches would deem underpowered, see Figure 1.

Only for a sample size of 500 participants or more would the majority of trials find a statistically significant difference in average low-density lipoprotein cholesterol between groups (Figure 1). Even at a sample size of 700 around 30% of trials would be expected to make the 'wrong' inference of not rejecting the null hypothesis. This is consistent with a priori analytic estimates of sample size to address the hypothesis.

To inform decision making using cost-effectiveness as the criterion, reducing the sample size has little effect on the conclusion of whether to fund, recommend and participate in TEXT ME, see Figure 2. For every simulation for each sample size the decision to adopt TEXT ME led to cost savings shown on the y-axis and gains to health, measured by QALYs shown on the x-axis.

A sample size of 100 or more in the primary trial would convince a risk neutral and rational decision maker that TEXT ME is both cost-saving and health improving, and so should be adopted. The imprecision surrounding this inference increases as the sample size reduces, but the decision-making inference does not change. If the goal is to make a good decision about whether TEXT ME should be adopted widely, then that could have been achieved with a much smaller trial, one that enrolled as few as 100 patients. This would have been a cheaper and quicker research project releasing scarce research dollars for other important projects.

When we simulated studies where there was no treatment effect, all the costs of implementing the TEXT ME program of around 1.5 million dollars for the cohort of 50,000 patients were incurred, but none of the health benefits and associated cost savings were realised. The estimates of change to health benefits straddled the zero line with a spread covering a relatively small change in QALYs of around 20 lost to 12 gained. The inference for decision makers is clear at any sample size that adoption would be a poor decision (Figure 3).

Dataset 1. Data used for a simulation of Figure 2

http://dx.doi.org/10.5256/f1000research.15522.d212377

Discussion

RCTs have become "massive bureaucratic and corporate enterprises, demanding costly infrastructure for research design, patient care, record keeping, ethical review, and statistical

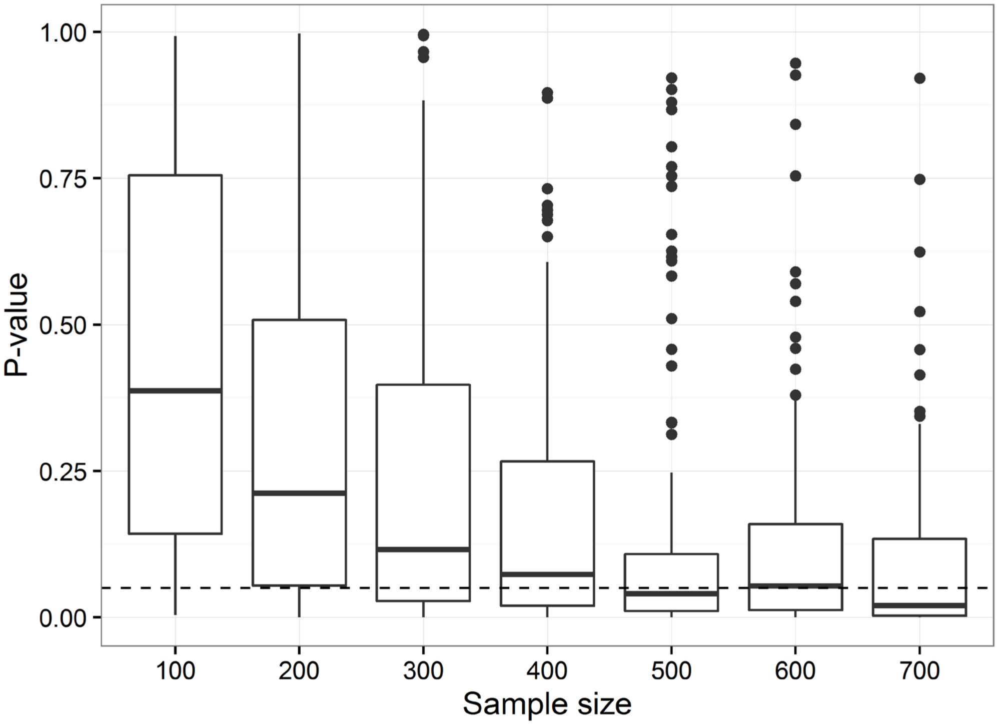

Figure 1. P-values increase as sample sizes decrease for the observed differences in low-density lipoprotein cholesterol (based on 500 simulations per sample size). The dotted horizontal line is the standard 5% threshold. The boxes are the 25th and 75th percentiles with the median as the central line. The upper whisker extends from the third quartile to the largest value no further than 1.5 * IQR from the quartile (where IQR is the inter-quartile range). The lower whisker extends from the 1st quartile to the smallest value at most 1.5 * IQR of the quartile. Data beyond the end of the whiskers are called 'outlying' points and are plotted individually.

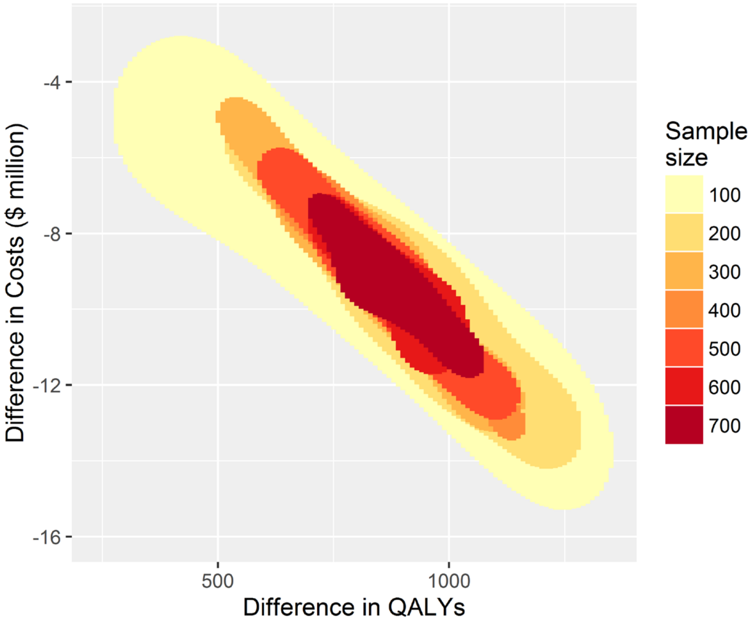

Figure 2. The conclusion for decision-making becomes more uncertain but does not change with decreasing sample size. The x-axis shows the QALY gains for TEXT ME over usual care, and the y-axis shows the cost savings.

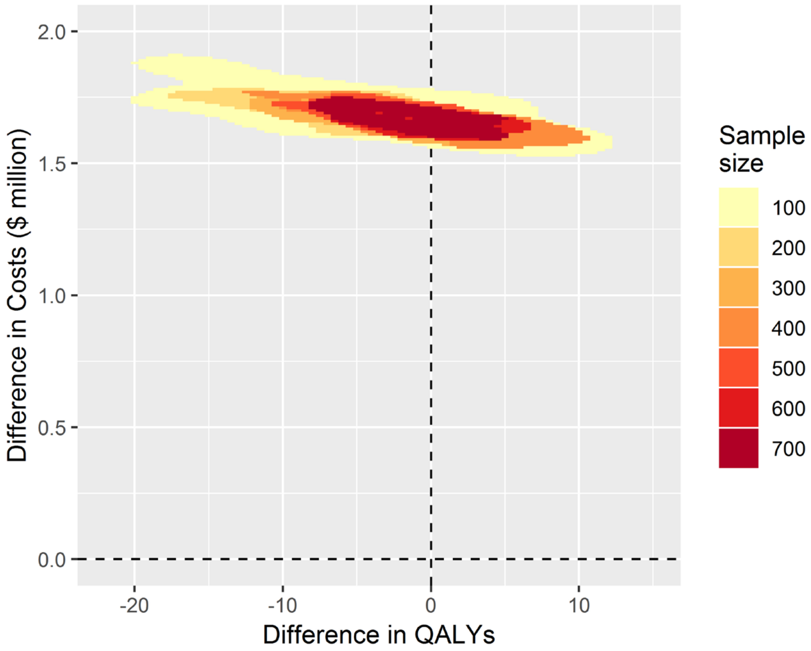

Figure 3. The conclusion for decision-making is clear when there is no treatment effect, costs are increased for no change to health benefits for all sample sizes.

analysis”[17]. A single phase 3 RCT could today cost $30 million or more[18] and take several years from inception to finalisation. These trials are powered for arbitrary rules of statistical significance. Critics of this approach[3] argue “that some of the sciences have made a mistake, by basing decisions on statistical significance” and that “in daily use it produces unchecked a loss of jobs, justice, profit, and even life”. The mistake made by the so called ‘sizeless scientist’ is to favour ‘Precision’ over ‘Oomph’. A ‘sizeless scientist’ is more interested in how precisely an outcome is estimated and less interested in the size of the implications for society or health services of any observed change in the outcome. They do not appear interested in the facts that “significant does not mean important and insignificant does not mean unimportant”. Even experts in statistics have been shown to interpret evidence poorly, based on whether the p-value crosses the threshold of 5% for statistical significance[19].

Researchers today are calling for a shift towards research designed for decision making[20]. Yet this is not new, in 1967 Schwartz & Lellouch[21] made a distinction between ‘explanatory’ and ‘pragmatic’ approaches. The former seeks ‘proof’ of the efficacy of a new treatment and the latter is about ‘choosing’ the best from two treatments. Patients, clinicians and payers of health care are interested in whether some novel treatment or health programme should be adopted over the alternatives.

There are many choices to be evaluated and many useful clinical trials to be undertaken, yet research budgets to support these are insufficient[22]. Funding a larger number of smaller trials to enable correct decisions about how to organise health services more frequently is a sensible goal. A hypothesis-testing approach maintains that a uniform level of certainty around these decisions is desirable, and needed by all stakeholders: managers, clinicians and patients. Yet the costs and benefits of every decision made are context-specific. Striving to eliminate uncertainty is likely to be inefficient use of research funding, where the benefit of achieving a given level of certainty is low or the prescribed precision unnecessary. We are not the only group that are advocating for this approach, and others have used cost-effectiveness as a criteria for dynamically deciding the necessary size of an ongoing trial[23]. There is a wider literature on decision making including economic data. Decision-making should address the costs and benefits throughout the life cycle of an intervention[24], with consideration of whether decisions could be made based on current evidence and whether additional research needs to be undertaken[25]. Other considerations for decision making under conditions of uncertainty have been established and reviewed in detail[26].

Our observations contradict advice by Nagendran *et al.*[27] who suggest researchers aim to “conduct studies that are larger and properly powered to detect modest effects”. This approach promotes using p-values for decision making without a more encompassing evaluation of all outcomes that are relevant for decision-making.

We suggest the decision making approach to sample size calculation would often lead to smaller trials, but not always. If rare adverse events had a substantial impact on cost and health outcomes the trial may be larger than a hypothesis testing trial powered for a single outcome, which was not the adverse event. This may especially be the case for trials of new drugs. There are some good arguments against smaller trials. A large trial with lots of data might help future proof an adoption decision. If costs, frequencies of adverse events or baseline risks change over time then a large trial might render sufficient information to defend the adoption decision in the future as compared to a small trial. There might also not be another opportunity to run an RCT, for ethical or funding reasons, and so gathering a lot of data when the chance arises could be wise. Smaller trials, despite being well designed, might find a positive result that overestimates the real effect[28]. This may have happened with our example of TEXT ME and a more conservative estimate of the intervention effect would likely come from a meta-analysis or repeated trial. Indeed Prasad *et al.*[29] found from 2,044 articles published over 10 years in a leading medical journal, 1,344 were about a medical practice, 363 of them tested an established medical practice and for 146 (40%) the finding was that practice was no better or worse than the comparator implying a reversal of practice. Those who deliver health services are unlikely to be rational and risk neutral. There is often scepticism and inertia when a change to practice is suggested and some clinicians will only change when evidence is overwhelming. Lau *et al.*[30] did a cumulative meta-analysis of intravenous streptokinase for acute myocardial infarction with mortality as the primary outcome. They showed the probability the treatment reduced mortality was greater than 97.5% by 1973 after 2,432 patients had been enrolled in eight trials. By 1977, after 4,084 patients had been enrolled in thirteen trials the probability the treatment was effective was more than 99.5%. By 1988, 36 trials had been completed with 36,974 patients included confirming the previous conclusion.

Our case study demonstrates - for a single carefully conducted trial - that more information might have been collected than was necessary for a good decision to be made about a decision to adopt the intervention. We did not cherry pick this trial, but selected it because it was a recent economic analysis and had broad implications for health. The differences in necessary sample sizes and evidence will depend on context and design of trials. It might often be that smaller and so faster and cheaper trials are sufficient for good decision-making. This would release scarce research dollars that funding bodies could use for other valuable projects. Our approach is part of the drive toward increasing the value of health and medical research, which currently has a poor return with an estimated 85% of investment wasted[31]. Further, as adaptive trials gain traction, decision based designs provide flexibility, facilitating faster evolution of implementable findings.

Data availability

The datasets used and/or analysed for the TEXT ME trial are not publicly available due to data sharing not being approved by the local ethics committee. To access the data, the corresponding author of the primary trial should be contacted (cchow@georgeinstitute.org.au).

A random sample of the TEXT ME clinical trial data that has similar features to the TEXT ME data is provided in the code used to create the simulations and figures, which is available on GitHub: https://github.com/agbarnett/trials.smaller

Archived code as at time of publication: http://doi.org/10.5281/zenodo.1322459[32]

Dataset 1: Data used for a simulation of Figure 2. DOI, 10.5256/f1000research.15522.d212377[33]

Grant information

The TEXT ME trial was supported by peer-reviewed grants from the National Heart Foundation of Australia Grant-in-Aid (G10S5110) and a BUPA Foundation Grant. We acknowledge the team who designed and conducted the TEXT ME trial and allowed us to re-analyse the data for the purpose of this paper: Clara Chow, Julie Redfern, Graham Hillis, Aravinda Thiagalingam, Stephen Jan, Maree Hackett, Robyn Whittaker. They did not provide editorial input or endorsement. The TEXT ME trial was administered by The George Institute for Global Health, Sydney Medical School, University of Sydney, Sydney, Australia.

Regarding the extra activity for this paper the authors declare that no grants were involved in supporting this work.

References

1. Hunink MM, Weinstein MC, Wittenberg E, *et al.*: **Decision making in health and medicine: integrating evidence and values.** Cambridge University Press; 2014.
Publisher Full Text
2. Tversky A, Kahneman D: **The framing of decisions and the psychology of choice.** *Science.* 1982; **211**(4481): 453–8.
PubMed Abstract | Publisher Full Text
3. Ziliak S, McCloskey D: **The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.** Ann Arbor, MI.: The University of Michigan Press; 2008.
Publisher Full Text
4. Woodcock J, Ware JH, Miller PW, *et al.*: **Clinical Trials Series.** *N Engl J Med.* 2016; **374**(22): 2167.
Publisher Full Text
5. Claxton K: **The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies.** *J Health Econ.* 1999; **18**(3): 341–64.
PubMed Abstract | Publisher Full Text
6. Goodman SN: **Toward evidence-based medical statistics. 1: The *P* value fallacy.** *Ann Intern Med.* 1999; **130**(12): 995–1004.
PubMed Abstract | Publisher Full Text
7. Chow CK, Redfern J, Hillis GS, *et al.*: **Effect of Lifestyle-Focused Text Messaging on Risk Factor Modification in Patients With Coronary Heart Disease: A Randomized Clinical Trial.** *JAMA.* 2015; **314**(12): 1255–63.
PubMed Abstract | Publisher Full Text
8. Wasserstein RL, Lazar NA: **The ASA's statement on *p*-values: context, process, and purpose.** *Am Stat.* 2016; **70**(2): 129–33.
Publisher Full Text
9. Claxton K, Palmer S, Longworth L, *et al.*: **A Comprehensive Algorithm for**

Approval of Health Technologies With, Without, or Only in Research: The Key Principles for Informing Coverage Decisions. *Value Health.* 2016; **19**(6): 885–91.
PubMed Abstract | Publisher Full Text

10. Phelps CE, Mushlin AI: **On the (near) equivalence of cost-effectiveness and cost-benefit analyses.** *Int J Technol Assess Health Care.* 1991; **7**(1): 12–21.
PubMed Abstract | Publisher Full Text

11. Torrance GW: **Measurement of health state utilities for economic appraisal.** *J Health Econ.* 1986; **5**(1): 1–30.
PubMed Abstract | Publisher Full Text

12. Asaria M, Griffin S, Cookson R: **Distributional Cost-Effectiveness Analysis: *A Tutorial*.** *Med Decis Making.* 2016; **36**(1): 8–19.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Claxton K: **Bayesian approaches to the value of information: implications for the regulation of new pharmaceuticals.** *Health Econ.* 1999; **8**(3): 269–74.
PubMed Abstract | Publisher Full Text

14. Redfern J, Thiagalingam A, Jan S, *et al.*: **Development of a set of mobile phone text messages designed for prevention of recurrent cardiovascular events.** *Eur J Prev Cardiol.* 2014; **21**(4): 492–9.
PubMed Abstract | Publisher Full Text

15. Chow CK, Redfern J, Thiagalingam A, *et al.*: **Design and rationale of the tobacco, exercise and diet messages (TEXT ME) trial of a text message-based intervention for ongoing prevention of cardiovascular disease in people with coronary disease: a randomised controlled trial protocol.** *BMJ Open.* 2012; **2**(1): e000606.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Burn E, Nghiem S, Jan S, *et al.*: **Cost-effectiveness of a text message programme for the prevention of recurrent cardiovascular events.** *Heart.* 2017; **103**(12): 893–4.
PubMed Abstract | Publisher Full Text

17. Bothwell LE, Greene JA, Podolsky SH, *et al.*: **Assessing the Gold Standard--Lessons from the History of RCTs.** *N Engl J Med.* 2016; **374**(22): 2175–81.
PubMed Abstract | Publisher Full Text

18. Sertkaya A, Birkenbach A, Berlind A, *et al.*: **Examination of clinical trial costs and barriers for drug development: report to the Assistant Secretary of Planning and Evaluation (ASPE)**. Washington, DC: : Department of Health and Human Services; 2014.
Reference Source

19. McShane BB, Gal D: **Statistical Significance and the Dichotomization of Evidence.** *J Am Stat Assoc.* 2017; **112**(519): 885–95.
Publisher Full Text

20. Lieu TA, Platt R: **Applied Research and Development in Health Care - Time for a Frameshift.** *N Engl J Med.* 2017; **376**(8): 710–3.
PubMed Abstract | Publisher Full Text

21. Schwartz D, Lellouch J: **Explanatory and pragmatic attitudes in therapeutical trials.** *J Clin Epidemiol.* 2009; **62**(5): 499–505.
PubMed Abstract | Publisher Full Text

22. Van Noorden R: **UK government warned over 'catastrophic' cuts.** *Nature.* 2010; **466**(7305): 420–1.
PubMed Abstract | Publisher Full Text

23. Pertile P, Forster M, La Torre D: **Optimal Bayesian sequential sampling rules for the economic evaluation of health technologies.** *J R Statist Soc A.* 2014; **177**(2): 419–438.
Publisher Full Text

24. Sculpher M, Drummond M, Buxton M: **The iterative use of economic evaluation as part of the process of health technology assessment.** *J Health Serv Res Policy.* 1997; **2**(1): 26–30.
PubMed Abstract | Publisher Full Text

25. Sculpher MJ, Claxton K, Drummond M, *et al.*: **Whither trial-based economic evaluation for health care decision making?** *Health Econ.* 2006; **15**(7): 677–87.
PubMed Abstract | Publisher Full Text

26. Claxton K, Palmer S, Longworth L, *et al.*: **Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development.** *Health Technol Assess.* 2012; **16**(46): 1–323.
PubMed Abstract | Publisher Full Text

27. Nagendran M, Pereira TV, Kiew G, *et al.*: **Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment.** *BMJ.* 2016; **355**: i5432.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Barnett AG, van der Pols JC, Dobson AJ: **Regression to the mean: what it is and how to deal with it.** *Int J Epidemiol.* 2005; **34**(1): 215–20.
PubMed Abstract | Publisher Full Text

29. Prasad V, Vandross A, Toomey C, *et al.*: **A decade of reversal: an analysis of 146 contradicted medical practices.** *Mayo Clin Proc.* 2013; **88**(8): 790–8.
PubMed Abstract | Publisher Full Text

30. Lau J, Schmid CH, Chalmers TC: **Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care.** *J Clin Epidemiol.* 1995; **48**(1): 45–57; discussion 59-60.
PubMed Abstract | Publisher Full Text

31. Chalmers I, Glasziou P: **Avoidable waste in the production and reporting of research evidence.** *Lancet.* 2009; **374**(9683): 86–9.
PubMed Abstract | Publisher Full Text

32. Barnett A: **agbarnett/smaller.trials: First release of R code for smaller clinical trials (Version v1.0).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1322459

33. Graves N, Barnett AG, Burn E, *et al.*: **Dataset 1 in: Smaller clinical trials for decision making; using p-values could be costly.** *F1000Research.* 2018.
http://www.doi.org/10.5256/f1000research.15522.d212377

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 2

Reviewer Report 11 October 2022

https://doi.org/10.5256/f1000research.17883.r151315

?

Daniel Benjamin Mark
Division of Cardiology, Department of Medicine, Duke University Medical Center, Durham, NC, USA

This article addresses an important and provocative idea, namely that we are making our clinical trials larger and more expensive than is necessary for rational decision making. I suspect the authors are correct but would offer several general reasons why I think the case they have presented overlooks some key elements that help explain why things are less than optimally efficient in the sense proposed by the authors.

First, the trial they use to illustrate their ideas is a relatively simple test of text messaging to improve risk factor management. The intervention was not expensive and did not involve any complex risk benefit considerations. However, most of the trials that affect clinical management and that have a large health care budgetary impact involve interventions that are very expensive (with the potential to generate annual expenditures in the US of multiple billions of dollars a year) and often involve complex issues of risk and benefits. Regulatory approval in the US has typically involved many smaller trials leading up to two pivotal phase III trials that generally both need to be "positive" on their primary endpoint, meaning they need to show the treatment effect has a p< 0.05. Failing to achieve the requisite level of "evidence" in this context will typically result in a decision not to grant regulatory approval, in which case the company developing the therapy is faced with the decision about whether to abandon the work and investment to that point or to invest more in what may be an even larger and more expensive next trial. Regulators in the US at least currently are charged with ensuring treatments are both safe and effective before granting market approval and they do not have any responsibility for addressing cost effectiveness or budget impact. So these decision makers at least would not accept a lower standard of evidence in order to improve efficiency and the level of evidence has a big effect on their decision making.

Second, clinical practice guideline committees and major clinical journals also triage clinical trial interpretation according to strict statistical significance criteria. Many trials published in journals such as NEJM have had the tested therapy declared ineffective because of a p value in the range of 0.06 to 0.10. In such situations, guideline committees are very likely to accept the official interpretation and make recommendations accordingly. These decision makers also do not have any interest in accepting lower levels of precision in order to improve the efficiency of the health care system. Part of the difficulty here is the (mis)understanding of what significance tests/p values can and cannot tell us about the outcomes of a clinical experiment/RCT, as the authors discuss. Fixing that by getting clinicians and statisticians and guideline committees to use more flexible, comprehensive interpretive approaches to evidence has proven quite difficult.

Third, the authors describe their target audience as health care decision makers managing health care services. The assumptions of economics regarding the decisions being made by rational decision makers who have the goal of maximizing the health benefits for the largest number of their population possible is an interesting model but it's not clear that it describes any actual health care system and how it functions. In the US, the primary focus is on budget impact not efficiency. In countries where economic analysis is required for reimbursement, budget impact still seems to be a dominant consideration.

The implications for the paper would seem to be that the authors should consider in their presentation a bit more about what the needed infrastructure is in order for their recommendations to be adopted. What kind of trials would be suitable for smaller tests accepting less precision in exchange for more rapid efficiency and lower costs? One area might be implementation trials, in other words, trials testing the deployment into the practice of things we already know work from prior, large pivotal trials. The example provided seems to fit in this category.

The global nature of scientific medicine means that the knowledge about effective but expensive therapies is widely available to providers and patients across health systems. Health systems can therefore no longer refuse to provide their citizens with advances simply because of the expense, but at the same time they have to control the growth of health care spending. Making more therapies available with a greater uncertainty accepted in the estimated treatment effect may be difficult to sell as a general concept. One could, I suppose, argue that if regulatory approval were not so expensive, companies would not need to charge such high prices for new advances. Not clear that there is any appetite among health regulators for moves in this direction and even if some countries did adopt such a program, the developers of new therapies facing the global market forces might not benefit enough to alter pricing.

Besides some forms of implementation research, I think what the authors are proposing could work in the context of "evidence free zones", areas of medicine where there is little beyond anecdote and expert opinion to use in decision making. In that context, an inexpensive trial that provides some reliable evidence, accepting a higher level of uncertainty, can still be used to change/guide practice and policy. Despite all the trials that get reported every year, much of current practice guidelines still consist of expert opinion, as large clinical trials are only done in select areas where funding sources exist willing to support the work. Most of clinical medicine falls outside these zones.

In summary, I think the authors have raised an important and interesting issue. I would ask them to consider discussing a bit more of the real world nuances of how and where this might work and where it is unlikely to be accepted.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical trials, outcomes research, clinical economics, cardiovascular medicine

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 09 September 2019

https://doi.org/10.5256/f1000research.17883.r46419

✓ **Steven A Julious**

Medical Statistics Group, University of Sheffield, Sheffield, UK

While I agree with the sentiments of the authors that there can be instances where a conventional sample size calculation may not be appropriate. I do have an issue with the generalisations they are making from a single trial to studies costing $30m.

The gold standard for an assessment of efficacy right or wrong is a formal hypothesis test. There is a need to definitively show evidence of clinical effect for most interventions.

The example that was quoted in the paper is a cheap intervention. Low cost interventions may be anticipated to observed small effects: so small that the expense of undertaking a clinical trial could be prohibitively expensive.

I have recent experience of two trials investigating low cost interventions in the U@UNI trial and the PLEASANT trial.[1,2,3,4] In both these trials the interventions could be shown to be cost effective (or cost saving). It could be contended in both the sample size could have been based on cost effectiveness.

In summary, therefore, there is merit in what the authors are suggesting but there needs to be consideration as to when the arguments could be applied.

References

1. Chloe T, Penny B, Mark S, Alan B, et al.: The cost-effectiveness of an updated theory-based online health behavior intervention for new university students: U@Uni2. *Journal of Public Health and Epidemiology*. 2016; **8** (10): 191-203 Publisher Full Text
2. Epton T, Norman P, Dadzie AS, Harris PR, et al.: A theory-based online health behaviour intervention for new university students (U@Uni): results from a randomised controlled trial.*BMC Public Health*. 2014; **14**: 563 PubMed Abstract | Publisher Full Text
3. Julious SA, Horspool MJ, Davis S, Bradburn M, et al.: PLEASANT: Preventing and Lessening Exacerbations of Asthma in School-age children Associated with a New Term - a cluster randomised controlled trial and economic evaluation.*Health Technol Assess*. **20** (93): 1-154 PubMed Abstract | Publisher Full Text
4. Franklin M, Davis S, Horspool M, Kua WS, et al.: Economic Evaluations Alongside Efficient Study Designs Using Large Observational Datasets: the PLEASANT Trial Case Study.*Pharmacoeconomics*. 2017; **35** (5): 561-573 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical statsitics, clinical trials

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 09 October 2018

https://doi.org/10.5256/f1000research.17883.r38819

?

Stephen Senn

[1] Competence Center for Methodology and Statistics, Luxembourg Institute Of Health, Strassen, Luxembourg

[2] School of Health and Related Research, University of Sheffield, Sheffield, UK

First let me apologise for a stupid slip in my first review. I referred to sampling *without* replacement and the authors quite rightly corrected me and pointed out that they sampled *with* replacement. That is, in fact, what I meant to say.

Second, let me acknowledge that the authors have gone some way to answering my criticisms. However, I am not completely satisfied, so before changing my overall judgement of the paper, I am going to explain the problem again.

Consider an alternative to sampling with replacement that will produce almost the same results they did. This is to compare a super trial in which the values for every patient in the TEXT ME trial are copied a huge number of times to create a million versions of each patient. Thus every patient in the trial has 999,999 identical virtual siblings. Where the TEXT ME trial had 710 patients we now have 710 million patients. Let us call this trial MegaText. If all these patients were real, then in the huge population of MegaText the effect seen would be the true effect. Now we can actually sample without replacement from MegaText and the result will be almost identical to sampling with replacement from the TEXT ME trial. The only slight difference is that in sampling without replacement from MegaText once a patient has been chosen there is a very slightly reduced chance of the patient being chosen again but since there are so many copies, this hardly matters.

Therefore, sampling from the TEXT ME trial is almost identical to sampling from the MegaText trial and it thus follows that we are sampling from a population in which there is a genuine treatment effect and there is no uncertainty about this. Thus the right decision is known to be to implement the program. Of course, and their simulation shows this, if the trial is small enough, you will sometimes choose the wrong treatment. If the purpose of the trial is pragmatic in the sense of Schwartz and Lellouch(Schwartz, D. & Lellouch, J., 1967), then it is this Type III error that has to be guarded against.

However, this raises the second issue. The decision to always choose what appears to be the better of two treatments being compared, however weak the evidence, is logical *if no further evidence can be obtained.* However, it is not necessarily logical if more information can be obtained at a modest cost. To see this, consider the case where a new treatment *N* is being compared to a standard treatment *S* and high values are good, as would be the case if we are measuring utility. The Bayesian posterior distribution for difference in effects (*N-S*) is mainly in the positive area: it is more probable than not, taking all things into account that it is better to use *N*. However, there is a non-negligible probability that actually S is better. If information can be obtained at low cost it may be worth doing so just to exclude the possibility that S is after all better.

Thus, I have some unease that the combination of starting with a proven treatment and showing that if one had carried out a smaller trial one would often come to an apparently good choice of treatment if one had to be made is quite as relevant to the practical problem that choosing a sample size is meant to solve. This is not to say that common approaches to doing this are good: far from it. They ignore the dimension of cost and this cannot be rationale.

Thus, if what the authors wish to say is: “just because a trial has not found a significant result it does not follow that it cannot be used to decide to implement a new treatment if no further information will be forthcoming” there are some circumstances under which I could agree. If they wish to imply that standards should generally be less stringent than they currently are, I do not think this sort of investigation is particularly relevant.

References

1. Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials.*J Chronic Dis*. 1967; **20** (8): 637-48 PubMed Abstract

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: As far as I am aware I have no competing interests. I maintain a general statement of interests here: http://www.senns.demon.co.uk/Declaration_Interest.htm

Reviewer Expertise: I am a medical statistician with many years experience in dealing with problems associated with drug development and regulation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 31 Oct 2019

Nicholas Graves, Queensland University of Technology, Brisbane, Australia

We thank the reviewer for the additional comments and try to respond to the key issue that arose in the first review.

The reviewer says "*simulating only from the case where the intervention is beneficial is not adequate*", but we simulate because we are not certain the intervention is beneficial. Hence, we simulate from the observed data in order to see whether the intervention is beneficial on a meaningful scale.

In general, these simulations use multiple outcomes, some of which may have a positive mean (e.g., improvement in blood pressure and the associated health benefits/ health utility) and some have a negative mean (e.g., increase in costs). Hence, it is a composite estimate that is more complex that just one mean being positive or not.

We used this approach for another important question about whether a hand hygiene campaign should be funded and showed the conclusions varied for different states and territories of Australia, see Table 3 of this paper...

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0148190

In South Australian, Tasmania & Western Australia the positive difference in the main outcome was not large enough to justify a decision to continue the campaign. But in Queensland, ACT and New South Wales we concluded the opposite.

We are not simply reliant on a single positive mean, which should happen for around 50% of studies where the intervention has no effect and hence is quite easy to achieve. Instead, we are looking at whether the observed difference is meaningful using the observed variation in the sample.

It is possible to use scenario analysis to simulate results under more pessimistic scenarios, such as a null treatment effect but reduction in costs, and these can be informative for decision makers.

Competing Interests: None

Version 1

Reviewer Report 07 September 2018

https://doi.org/10.5256/f1000research.16927.r37678

?

Stephen Senn

[1] Competence Center for Methodology and Statistics, Luxembourg Institute Of Health, Strassen, Luxembourg

[2] School of Health and Related Research, University of Sheffield, Sheffield, UK

The authors propose that when a clinical trial is sought to inform practical decision-making, conventional standards of 'proof' may be too stringent and in consequence resources may be wasted. They illustrate this by simulating from a particular clinical trial, the TEXT ME trial, using progressively smaller sample sizes and suggest that a useful decision could have been made with fewer patients.

The general argument presented is interesting and the conclusion that trials are sometimes too big if practical decision making is the object may well be correct. In this respect, a key distinction was made just over 50 years ago by Schwartz and Lelouch[1] between what they called *explanatory* or *pragmatic* approaches. In the former case 'proof' of the efficacy of a new treatment may be sought. In the latter case one may simply wish to choose the (plausibly) better of two treatments.

However, unless I have misunderstood what the authors are doing (which I do not exclude but in that case they should clarify this) the simulation is not a valid proof of what they claim, even for the example chosen.

The problem is the following. By simulating from the particular trial results, they are simulating from a universe in which the treatment is effective. This would be true even if the results from the TEXT ME trial had not been 'significant'. It is true of any trial in which the observed results favour the intervention. To see this consider that valid statistical analyses will typically have type I error rates in excess of a chosen nominal value if the mean under the intervention is greater (assuming high values are good) than the mean in the control group in the population in question. Provided that the type I error rate is controlled when this is not the case, this is a desirable property of such tests.

Usually, the population in question is taken to be the population of all possible randomisations of the patients. Here, the authors sampled without replacement from the population. The population from which they are sampling is the population of results in the full TEXT ME trial. However, this is a population in which on average the results were better for the intervention.

Hindsight is an exact science but those making practical healthcare decisions are involved in the quite different game of foresight and they need to know whether the decision they are about to make is a reasonable one. This requires their allowing for the possibility that the intervention is useless or even harmful. Thus a mixture of possible situations has to be considered: simulating only from the case where the intervention is beneficial is not adequate.

In fact the precise nature of the mixture envisaged can have a huge effect on the inferences. Recently, a number of authors have called for statistical standards of evidence to be modified in the opposite direction. For instance Benjamin *et al.*[2] have suggested that the standard of p=0.005 should be adopted. David Colquhoun[3] has proposed an even more stringent standard of P=0.001. This flows from the particular approach to Bayesian hypothesis testing which places a lump of probability on no difference between treatments. (See my blog[4] for a discussion.) In my opinion, these are not good suggestions for a number of reasons, including that such prior distributions are far too informative and that these authors implicitly assume, which is far from obviously the case, that the explanatory purpose of clinical trials is more important than the pragmatic one.

However, I agree entirely with the authors, that as soon as practical decision-making involving economics is involved, it is the value of information that is important. In this connection, I can recommend the work of Forster, Pertile and colleagues[5,6]. See also Burman *et al*. [7]

Thus, I think to make good their claim, the authors would, at the very least, need to simulate from a universe in which the intervention was not necessarily better than the control. Unless I have misunderstood, this was not the simulation they undertook.

References

1. Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials.*J Chronic Dis*. 1967; **20** (8): 637-48 PubMed Abstract
2. Benjamin D, Berger J, Johannesson M, Nosek B, et al.: Redefine statistical significance. *Nature Human Behaviour*. 2018; **2** (1): 6-10 Publisher Full Text
3. Colquhoun D: An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci*. 2014; **1** (3): 140216 PubMed Abstract | Publisher Full Text
4. Senn SJ: Double Jeopardy: Judge Jeffreys upholds the law. 2015. Reference Source
5. Pertile P, Forster M, Torre D: Optimal Bayesian sequential sampling rules for the economic evaluation of health technologies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2014; **177** (2): 419-438 Publisher Full Text
6. Jobjörnsson S, Forster M, Pertile P, Burman CF: Late-stage pharmaceutical R&D and pricing policies under two-stage regulation.*J Health Econ*. 2016; **50**: 298-311 PubMed Abstract | Publisher Full Text
7. Burman C-F: Decision Analysis in Drug Development. In: Dmitrienko A, Chuang-Stein C, Agostino R, eds. Pharmaceutical Statistics Using SAS: A Practical Guide.*Cary: SAS Institute*. 2007. 385-428
8.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am a medical statistician with many years experience in dealing with problems associated with drug development and regulation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Sep 2018

Nicholas Graves, Duke NUS Graduate Medical School, Singapore

The authors propose that when a clinical trial is sought to inform practical decision-making, conventional standards of 'proof' may be too stringent and in consequence resources may be wasted. They illustrate this by simulating from a particular clinical trial, the TEXT ME trial, using progressively smaller sample sizes and suggest that a useful decision could have been made with fewer patients.
The general argument presented is interesting and the conclusion that trials are sometimes too big if practical decision making is the object may well be correct. In this respect, a key distinction was made just over 50 years ago by Schwartz and Lelouch1 between what they called explanatory or pragmatic approaches. In the former case 'proof' of the efficacy of a new treatment may be sought. In the latter case one may simply wish to choose the (plausibly) better of two treatments.
RESPONSE: Thanks for flagging this interesting paper on trials and pragmatic decision making. We certainly agree with them that, “many trials would be better approached pragmatically.” We have included this paper in the discussion section.
However, unless I have misunderstood what the authors are doing (which I do not exclude but in that case they should clarify this) the simulation is not a valid proof of what they claim, even for the example chosen.
RESPONSE: Our aim was to illustrate the principles of this approach using a case study rather than provide “proof” that this approach is always better. We have changed the title to reflect this.
The problem is the following. By simulating from the particular trial results, they are simulating from a universe in which the treatment is effective. This would be true even if the results from the TEXT ME trial had not been 'significant'. It is true of any trial in which the observed results favour the intervention. To see this consider that valid statistical analyses will typically have type I error rates in excess of a chosen nominal value if the mean under the intervention is greater (assuming high values are good) than the mean in the control group in the population in question. Provided that the type I error rate is controlled when this is not the case, this is a desirable property of such tests.
RESPONSE: We agree, although our approach includes the changes to costs from implementing the TEXT ME intervention, so the mean difference also has to also be

practically significant in order to recover these costs.

Usually, the population in question is taken to be the population of all possible randomisations of the patients. Here, the authors sampled without replacement from the population. The population from which they are sampling is the population of results in the full TEXT ME trial. However, this is a population in which on average the results were better for the intervention.

RESPONSE: We sampled with replacement. Using our approach, the group means were not always greater in the intervention group. For the primary outcome of LDL cholesterol, the mean was worse in the TEXT ME sample compared with usual care for around 22% of simulations when using the smallest sample size of 100. The mean difference in the secondary outcome of systolic blood pressure was stronger in the original data, and in simulations the mean in the TEXT ME sample was always lower (better) compared with the usual care group.

Hindsight is an exact science but those making practical healthcare decisions are involved in the quite different game of foresight and they need to know whether the decision they are about to make is a reasonable one. This requires their allowing for the possibility that the intervention is useless or even harmful. Thus a mixture of possible situations has to be considered: simulating only from the case where the intervention is beneficial is not adequate.

RESPONSE: Our aim is to provide results that are useful for decision makers, including estimates of uncertainty about the decision.

In fact the precise nature of the mixture envisaged can have a huge effect on the inferences. Recently, a number of authors have called for statistical standards of evidence to be modified in the opposite direction. For instance Benjamin et al.2 have suggested that the standard of p=0.005 should be adopted. David Colquhoun3 has proposed an even more stringent standard of P=0.001. This flows from the particular approach to Bayesian hypothesis testing which places a lump of probability on no difference between treatments. (See my blog4 for a discussion.) In my opinion, these are not good suggestions for a number of reasons, including that such prior distributions are far too informative and that these authors implicitly assume, which is far from obviously the case, that the explanatory purpose of clinical trials is more important than the pragmatic one.

RESPONSE: We agree and the tension of the explanatory versus pragmatic trial is a key motivation for this paper. These adjustments to the use of the p-value remain focused on the p-value and it how can inform decisions. These adjustments have been motivated by prior abuses and misinterpretations of the p-value, which is a prosaic statistic. Our approach aims to give decision makers, working under conditions of scarce resources, more meaningful statistics regarding changes to costs and health benefits.

However, I agree entirely with the authors, that as soon as practical decision-making involving economics is involved, it is the value of information that is important. In this connection, I can recommend the work of Forster, Pertile and colleagues5,6. See also Burman et al. 7

RESPONSE: Thanks for flagging these interesting papers.

Thus, I think to make good their claim, the authors would, at the very least, need to simulate from a universe in which the intervention was not necessarily better than the control. Unless I have misunderstood, this was not the simulation they undertook.

RESPONSE: We have added just such a simulation, which shows that when there’s no treatment benefit there is a positive cost from the intervention that is not outweighed by any quality of life benefit. The cost-effectiveness plot shows clear evidence against adopting the intervention. We have added the methods and results for this new simulation and include Figure 3.

References

1. Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials.J Chronic Dis. 1967; 20 (8): 637-48 PubMed Abstract
2. Benjamin D, Berger J, Johannesson M, Nosek B, Wagenmakers E, Berk R, Bollen K, Brembs B, Brown L, Camerer C, Cesarini D, Chambers C, Clyde M, Cook T, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field A, Forster M, George E, Gonzalez R, Goodman S, Green E, Green D, Greenwald A, Hadfield J, Hedges L, Held L, Hua Ho T, Hoijtink H, Hruschka D, Imai K, Imbens G, Ioannidis J, Jeon M, Jones J, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell S, McCarthy M, Moore D, Morgan S, Munafó M, Nakagawa S, Nyhan B, Parker T, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt F, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts D, Winship C, Wolpert R, Xie Y, Young C, Zinman J, Johnson V: Redefine statistical significance. Nature Human Behaviour. 2018; 2 (1): 6-10 Publisher Full Text
3. Colquhoun D: An investigation of the false discovery rate and the misinterpretation of p-values.R Soc Open Sci. 2014; 1 (3): 140216 PubMed Abstract | Publisher Full Text
4. Senn SJ: Double Jeopardy: Judge Jeffreys upholds the law. 2015. Reference Source
5. Pertile P, Forster M, Torre D: Optimal Bayesian sequential sampling rules for the economic evaluation of health technologies. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2014; 177 (2): 419-438 Publisher Full Text
6. Jobjörnsson S, Forster M, Pertile P, Burman CF: Late-stage pharmaceutical R&D and pricing policies under two-stage regulation. J Health Econ. 2016; 50: 298-311 PubMed Abstract | Publisher Full Text
7. Burman C-F: Decision Analysis in Drug Development. In: Dmitrienko A, Chuang-Stein C, Agostino R, eds. Pharmaceutical Statistics Using SAS: A Practical Guide.Cary: SAS Institute. 2007. 385-428

Competing Interests: None