Reviewer's report

Title: Online Randomized Controlled Experiments at Scale: Lessons and Extensions to Medicine

Version: 0 Date: 13 Jun 2019

Reviewer: Sarah Lensen

Reviewer's report:

Thank-you for the opportunity to review this manuscript which describes the use of A/B testing (randomisation) in a technology industry, and touches on how medicine might learn from the scale and experiences of businesses using this technology, such as Google.

Overall I think this article presents an interesting concept, which readers may not be familiar with and would be of interest to them. It is authored by those working directly with this technology in industry, who have many valuable insights. My main concern is the possible lack of application - or at least lack of explanation of possible application - to medicine. Without clear ways to implement this technology in medicine readership may be reduced and therefore the long length of the article may not be justified.

I have the following more specific comments and suggestions:

1. The article is very long and some areas feel repetitive, and lacking in value-add. The title leads the reader to believe the article will discuss the (potential) application of A/B testing in medicine, however the article largely focusses on the experiences of the involved companies with little explanation or example of how these features could be applied in medicine. Particularly the sections under the subheadings 'Culture change and evolution of organizational processes' and 'Experimentation maturity on multiple axes' are too long and may not add value, for example the discussion about the MVP is probably not necessary.

2. I find the section 'Some additional technical issues' is so general that it is not useful.

3. The example of applying A/B testing in AVS is given, which is very helpful.

   a. I have not heard this term before (AVS), and I wonder if this might be common in the USA only. Perhaps you could briefly expand on what an AVS is (a written summary of a healthcare interaction which is taken home with or otherwise distributed to the patient after a visit), with reference to the US. Similarly the HIPAA will not be familiar to non-US readership.

   b. The application of A/B testing in medicine is lacking in terms of both the intervention being tested and the outcome being measured.

   i. Intervention: one of the scenarios posed relates to the delivery channel of the AVS - but can A/B testing really compare postal letter vs mobile delivery? I think bullet point 5 - regarding information presentation in the AVS - is of most applicability, and could be expanded, for example to explain how changes to the app delivering the AVS can be modified.
ii. Outcome: these companies use A/B testing to modify their apps etc to increase revenue. Revenue is both their motivation to conduct A/B testing and the key outcome measured. In healthcare, what can we measure as a consequence of changes to the app appearance? Perhaps 1) whether patients read the complete AVS, 2) whether they return to open the AVS at a later date - which might be a surrogate for adherence to the medical advice contained etc. It might assist the reader to understand what exactly our motivation would be for conducting A/B in medicine? Ultimately patient health? Later on in the paper the potential linkage to electronic medical records is discussed but only briefly - this must be where this testing would really add value, if we could measure the actual health outcomes for those randomised.

c. After the AVS example is provided, there is very little further reference or extrapolation of the A/B testing experiences of Google etc. to the healthcare setting. The remaining sections of the manuscript might benefit from continued analogy with AVS or other possible examples of using A/B testing in medicine - currently it is largely focused on company revenue and whether the user 'pays' for something.

d. Are there other examples A/B testing could be used for other than just AVS? In the UK, patients are able to book GP visits on a mobile app; could A/B testing be useful in this setting to test the best methods to ensure patients click through and secure a booking? I feel these sorts of examples are lacking.

4. An example of the unclear transferability to healthcare is highlighted in the discussion of power. It is noted that the effect sizes anticipated in A/B testing are usually small (0.5%), much smaller than we would dream of testing for in an RCT, both because the sample size of the required RCT is infeasibly large, and because very small effect sizes are not normally deemed clinically meaningful. Therefore, it might be helpful to expand on why such small differences in the hands of A/B testing remain relevant to medicine. It is stated that a 0.5% effect might cost Google >10 million dollars, but where is the analogy for medicine? We don't have the number of Google users using any given healthcare app so the translation of 0.5% effect to a population level or any actual impact on health is unclear.

5. I did not understand the example given regarding "In the medical field, such short time effects might be observed in a perioperative setting with duration of hospital stay, hospital mortality, complications or 30-day re-admission rates."

6. I wonder if it might be in scope to discuss the potential ethical concerns in transferring this testing to routine medicine, both in terms of

a. Ethics approval/oversight - normally when we test any intervention related to healthcare we seek ethics approval, even for interventions which are service management delivery and may be cluster-randomised. For example, an RCT comparing different pamphlets for cancer screening uptake would need ethics approval. Would this be necessary for A/B testing of an app to deliver AVS?

b. The Google-consumer relationship is a rather different scenario to the Dr-patient or Hospital-patient relationship. How would patients feel about being exposed to the testing of the delivery of their health information via AVS? It is conceivable that, especially moving towards anything like the "test everything with controlled experiments" philosophy, that certain experiments may result in the patients not receiving information that is important to their healthcare. You mentioned the bugs of overlapping
experiments resulting in the 'buy now button' disappearing below the end of the page. If such a bug arose in the case of an AVS, and important specific healthcare instructions were pushed off the bottom of the screen and not seen by patients, this may be viewed as a more problematic consequence than Amazon losing the sale of a handbag. Also - the company only realises the bug has occurred because the key outcome of interest is measurable within the system (revenue). Whereas, healthcare providers would remain unaware that their patient has not read the full instructions and may be at risk.

7. I notice the paper is submitted as a methodology piece, and I wonder if it might be better as a commentary. It does not introduce new methodology per se, but discusses the theoretical application of A/B testing methodology in a new area

**Level of interest**
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable

**Quality of figures**

All images and figures within the manuscript should be genuine i.e. without evidence of manipulation. No specific feature within an image may be enhanced, obscured, moved, removed, or introduced. If you have concerns about the veracity of the figures you should choose the first option below.

**Statistical review**

Is it essential that this manuscript is seen by an expert statistician? If so, please give your reasons in your report.
**Declaration of competing interests**

Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license ([http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/)). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal

Were you mentored through this peer review?

No