Author’s response to reviews

Title: Key challenges for delivering clinical impact with artificial intelligence

Authors:

Christopher Kelly (cjkelly@google.com)

Alan Karthikesalingam (alankarthi@google.com)

Mustafa Suleyman (mustafasul@google.com)

Greg Corrado (gcorrado@google.com)

Dominic King (dominicking@google.com)

Version: 1 Date: 29 Aug 2019

Author’s response to reviews:

Dear Dr Lopez Munoz,

Thank you very much for the opportunity to revise the manuscript “Key challenges for delivering clinical impact with artificial intelligence” (BMED-D-19-00867).

Please find attached our responses below.

Reviewer #1 - John P. Fox, PhD

This paper is one of a clutch of recent overviews of AI in medicine (to which this reviewer has contributed) but it is particularly important because it is provided by a highly influential organisational that has led key developments in AI/machine learning (Google Deepmind). While it is a technically well informed and sophisticated discussion the authors are also mostly medically trained and include senior members of the company (its CMO and one of its founders).

Despite the authors' having a commercial interest in the subject of the review the authors are very clear about the level of hype about ML alongside the current lack of evidence of adoption and clinical value (with the likely exception of medical imaging) and a variety of other significant challenges. This review is a valuable contribution that will help BMC Medicine readers and the wider healthcare community, medtech developers, AI researchers etc to have an informed conversation about what is required to deliver the potential benefits of AI for better patient care and improved clinical services.

>> Thank you very much for these comments.

This reviewer's most general concern is the absence of even a brief acknowledgement of the long history of research on using traditional symbolic AI techniques in medicine, which have also tried to
address some of the challenges that are rightly identified (e.g. the importance of explanation) and can be applied to others. In the following I will refer to the traditional approach as AI/KE (Knowledge Engineering). However this reviewer has worked on AI/KE for many years, so one might say I have a vested interest in promoting them so it seems appropriate to identify myself; my name is John Fox and my interests and expertise can be reviewed at linkedIn profile is https://www.linkedin.com/in/john-fox-914a19b/.

>> Thank you for highlighting this area. The manuscript did contain a brief discussion about why deep learning models struggle with explainability in part due to a lack of explicit knowledge representation - e.g.: under “Human barriers to AI adoption in healthcare”, page 6:

“Best performing models (e.g. deep learning) are often the least explainable, while models with poorer performance (e.g. decision trees) are the most explainable. A key current limitation of deep learning models is that they have no explicit declarative knowledge representation, leading them to have considerable difficulty in generating the required explanation structures [80].”

>> We have expanded upon this to add an acknowledgement to the long history of research in this field:

“Machine learning methods that build upon a long history of research in traditional symbolic AI techniques to allow for encoding of semantics of data and allow for using ontologies to guide the learning process may allow human experts to understand and retrace decision processes more effectively (Fox 1984; Lacave and Diez 2002).”

Specific comments

The paper starts with a very useful compilation of published trials and evaluation studies on AI/ML and an honest recognition of the limited adoption of these methods to date. It is commonly said that medical imaging seems closest to large scale adoption but this is only a small part of medical technology and an even smaller part of everyday clinical practice. AI/KE has not been extensively applied to imaging or signal processing generally, but it has significant progress in other aspects of routine care (e.g. many kinds of clinical decision-making, treatment planning, workflow management) so perhaps we should be looking for hybrid approaches?

>> We agree that knowledge engineering and expert systems have not been extensively applied to imaging, perhaps as the data are less amenable to this form of analysis. There are some promising examples - i.e. using transfer representation learning in otitis media (Chuen-Kai Shie et al. 2015), but given this is a small area of the field at present, we have opted not to include this in the review at this time.

I strongly support the authors' positions on rigorously designed and executed trials, the issue of prospective/retrospective comparisons etc - educating the technology community whose members are frequently new to medical research norms is key to clinical credibility. The start-up mantra "move fast and break things" is not appropriate in healthcare, where the primary goal is to maximise quality, safety and improved patient experience.

>> Thank you and we appreciate your support on this position.
Transparency, accountability, bias and ethics of AI are already topics of significant public concern. The suggestion that we need a better understanding of interaction between human and algorithm is correct. I would add that KE methods already address these issues to some extent and research on explainable AI by behavioural and cognitive scientists and others is increasingly focused on topics like meaningful representation of medical knowledge and expertise, explanation strategies and dialogues etc. This again speaks to the need for hybrid methods perhaps.

>> We agree that meaningful representation of medical knowledge and expertise, explanation strategies and dialogues will be potentially important ingredients to achieve human-computer interaction. We have expanded the relevant section to include this (“Developing a better understanding of interaction between human and algorithm”, page 7):

“It has also been shown that humans assisted by AI performed better than either alone in a study of diabetic retinopathy screening [84], and also in detection of cancer in lymph node histology (Wang et al. 2016). Techniques to more meaningfully represent medical knowledge, provide explanation, and facilitate improved interaction with clinicians will only improve this performance further. We need to continue gaining a better understanding of the complex and evolving relationship between clinicians and human centred AI tools in the live clinical environment [85].”

Over-generalisation across different patient populations, case mixes, territories etc is a well known concern (the pioneer of Bayesian diagnosis methods Tim de Dombal made this point in the seventies) and is exacerbated by a lack of transparency. In my view AI/KE offers a way of empowering clinicians to participate critically in design, development and deployment of point of care services like decision support, workflow and patient management services.

>> Thank you for this suggestion. We have amended the paragraph where this is discussed (“Algorithmic bias”, page 5) to read:

“A greater awareness of these issues and empowering clinicians to participate critically in system design and development will help guide researchers to ensure that the correct steps are taken to quantify bias before deploying models.”

I believe that the point about the ability to compare different algorithms in a way that is clinically intelligible is key to large scale engagement with clinicians, many of whom are currently bemused about whether AI is their friend or foe. I have recently completed a paper on this topic and would be happy to share and discuss it with the authors.

>> Thank you for this, and we agree that this approach will be key. We would be very interested to read the paper mentioned.

Author declarations

The authors say they are employed by Google LLC but that funding is "not applicable" - is that saying there is no conflict of interest? If that is the intention I think it needs explanation.

>> Our intention was to state that no specific funding (grant or otherwise) was provided for this review. However, all authors are employed by Google LLC, so we have amended this accordingly.
Reviewer #3 - Seongho Park

This manuscript is timely and addresses an important issue regarding AI in healthcare. Although OP/ED articles that overlap with the contents of this manuscript have appeared in some other journal, even including premier journals such as Lancet, JAMA, and Science, this manuscript is more comprehensive. Authors successfully listed virtually all relevant points and explained them with proper citation. The content organization is a bit different from what it would have been from a typical health technology assessment perspective. However, the current organization is also clear enough to follow. Nevertheless, some minor revisions considering the following comments and additional citations would be beneficial.

>> Thank you very much for these positive comments.

1. Since the recent boom of deep learning, AI and deep learning are often used almost synonymously. Therefore, it appears that the example studies quoted are deep learning related. Regarding RCT on AI, there is a large RCT named INFANT trial, "The INFANT Collaborative Group. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. Lancet. 2017; 389: 1719-1729." It is not deep learning technology but belongs to the category of AI. To my knowledge, this is the largest RCT of this kind. It is an excellent example of RCT on AI and also serves as an example that high accuracy may not guarantee a better patient outcome.

>> This is an excellent study, and we are pleased to include it in the paper. It serves as an important cautionary tale that higher accuracy delivered by automated systems does not necessarily result in better outcomes for patients. We were impressed that the CTG interpretation system used was described as early as 1994, and evaluated against experts in 1995 (Keith et al. 1995), combining various systems including signal processing, rule based algorithms and a small neural network.

The section on peer-reviewed RCTs has therefore been amended to read:

“These include an algorithm to detect childhood cataracts with promising performance in a small prospective study [54] but less accurate performance compared to senior clinicians in a diagnostic RCT [56]; a single blind RCT that showed a significantly reduced blind spot rate in esophagogastroduodenoscopy [57]; an open, non-blinded randomised trial of an automatic polyp detection algorithm for diagnostic colonoscopy demonstrating a significant increase in detection of diminutive adenomas and hyperplastic polyps [58]; and a simulated prospective, double-blind RCT of an algorithm to detect acute neurologic events [59][60]. The final study is a cautionary example of how higher accuracy enabled by AI systems does not necessarily result in better patient outcomes [60]. Future studies should aim to use clinical outcomes as trial endpoints to demonstrate longer-term benefit, while recognising that algorithms are likely to result in change of sociocultural context or the care pathway. This may necessitate more sophisticated approaches to evaluation [61].”

2. Regarding the quality reporting of AI studies, the mention of TRIPOD is appropriate. However, TRIPOD was made principally considering traditional statistical regression analysis in mind. TRIPOD statement specific to machine learning (TRIPOD-ML) is currently under development (Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019 Apr 20;393:1577-1579). Authors might want to address this briefly to make readers pay attention.
Thank you for this helpful suggestion - this should be included, and has now been added as a statement following the discussion of TRIPOD:

“In addition, a new version of the TRIPOD statement that is specific to machine learning prediction algorithms (TRIPOD-ML) is in development and will focus on the introduction of machine learning prediction algorithms, and establish methodological and reporting standards for machine learning studies in healthcare [63].”

3. Indeed, AUC may not necessarily represent clinical efficacy appropriately. Also, AUC itself does not work by itself for clinical decision making. It has to be accompanied by a threshold that turns the continuous AI output into decision categories. The same factors that limit the generalizability of AI performance would also prohibit the generalization of the threshold. For example, in Hwang et al. JAMA Netw Open. 2019;2:e191095 (authors' reference #5), the same high-sensitivity cutoff yielded a wide range of specificity results (0.566 to 1) across different institutions.

This is a useful point to add to the paper. We have improved the section on AUC (“Metrics often do not reflect clinical applicability”, page 3) to more explicitly include reference to model operating points that are required to convert a continuous model score into discrete decision categories:

“As well as reporting sensitivity and specificity at a selected model operating point (required to turn the continuous model output into discrete decision categories), papers should include information about positive and negative predictive value. As no single measure captures all desirable properties of a model, several measures are typically reported to summarise performance of a model. However, none of these measures ultimately reflect what is most important to patients - i.e. whether the use of the model results in a beneficial change in a patient’s care [66].”

We have also added a reference to decision curve analysis (i.e. certain ROC curve profiles may offer better patient benefit despite having a lower ROC AUC):

“Clinicians need to be able to understand how proposed algorithms could improve care of their patients within a relatable workflow, but most papers do not attempt to present such information. Potential approaches have been suggested including decision curve analysis, which aims to quantify the net benefit of using a model to guide subsequent actions [67].”

Finally, your point about the difficulty of generalisation of operating points is important to add, and your concrete example with Hwang et al is very useful. This seems most relevant in the generalisation section (p5, “Challenges in generalisation to new populations and settings”), added via with the following new statement:

“Generalisation of model operating points may also prove challenging across new populations, as illustrated in a recent study to detect abnormal chest radiographs, where specificity at a fixed operating point varied widely from 0.566-1.000 across five independent datasets [5].”

4. The importance of external validation using a representative sample of the intended deployment population in clinical practice is addressed in several locations in the manuscript. However, it would need to be mentioned more directly for emphasis as the lack of proper external validation of AI is currently a critical concern. According to a recent study (Kim et al. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical
Images: Results from Recently Published Papers. Korean J Radiol. 2019 Mar;20:405-410. https://doi.org/10.3348/kjr.2019.0025), which analyzed published studies investigating the performance of AI algorithms that analyze medical images to provide diagnostic decisions, only 6% performed external validation. Furthermore, none performed external validation using a prospective dataset representative of the intended clinical cohort.

>> Thank you - we agree that this should be called out more explicitly. We have added a new paragraph under “Challenges in generalisation to new populations and settings” (p5):

“Proper assessment of real world clinical performance and generalisation requires appropriately designed external validation. This involves testing an AI system using adequately sized datasets that are collected from institutions other than those that provided data for model training in order to adequately represent all relevant variations in patient demographics and disease states of target patients in real world clinical settings where the system will be applied [73]. This practice is currently rare in the literature and is of critical concern. A recent systematic review of studies that evaluated AI algorithms for the diagnostic analysis of medical imaging found that only 6% of 516 eligible published studies performed external validation [74].”

5. I recommend deleting the mention of economic incentives including 1) lines 52-53 on page 6, 2) the entire subsection titled "Misaligned incentives for adoption at scale" on page 7, and 3) "economic incentives to promote value-based care" in the last sentence of the conclusion paragraph). Unlike other issues, this issue is dependent upon the social and health system of each country. So, it is not generalizable knowledge. Furthermore, value-based healthcare and cost-effectiveness of a health technology need to take a lot more factors into account than those briefly mentioned in the manuscript. As the authors address in the manuscript, we have yet to investigate the real-world accuracy of AI and health impact of AI. It is yet premature to discuss the cost-effectiveness of AI.

>> This is a thoughtful suggestion and we agree that these points may not be necessary for this article, and the themes are already adequately addressed earlier in the manuscript. We have deleted suggestion (1), (2) and (3).

6. The conclusion paragraph should mention the importance of prospective studies and the importance of clinical trials that directly assess the clinical impact of AI beyond technical accuracy, i.e., how AI affects the quality of care, variability between healthcare professionals, efficiency/productivity of clinical practice, and most importantly patient outcomes. These are mentioned in the main body but is not clearly stated in the conclusion paragraph.

>> Thank you for this suggestion - we have adjusted the conclusion accordingly to incorporate this. Overall thank you very much for the excellent suggestions - we very much appreciate your input and feel the manuscript is significantly improved as a result.

We hope that following these revisions our manuscript is now ready for publication and we look forward to hearing from you.

Yours sincerely,

Christopher Kelly