Abstract for HINF Seminar Series: Jeremy Wyatt - Dec. 4, 2024 10AM PT

Improving the impact of AI in practice: we must go beyond testing performance to user impact studies, not straight to field trials

Speaker

Dr Jeremy C Wyatt FACMI FRCP FBCS, Emeritus Professor of Digital Healthcare, University of Southampton, Southampton, UK

Abstract

Despite huge investment in and political support for AI in healthcare, less than 10% of AI decision support tools are studied against a gold standard or outside the lab, so 91% have no impact on clinical practice [1]. Some do not even move beyond limited internal validation studies to check their accuracy on datasets from outside the development site. However, relying on the results of validation studies assumes that AI will be the decision taker but the reality is that AI is used to support humans taking decisions [2]. We also know that clinicians value evidence about the impact of the AI on user decisions more than validation study results [3, 4].

Most developers falsely assume that the only step beyond validating their model is a randomised controlled trial (RCT) of AI impact on real users and their decisions. While very useful, these raise many logistical and resource issues. For example, before an RCT is carried out an AI must be connected to the organisation's EPR, which means matching up clinical codes and passing cyber security checks, an intricate and risk-prone process requiring high-level signoff [5]. Designing a decision support RCT means finding a methodologist familiar with specific biases such as the carryover, checklist and Hawthorne effects and automation bias [6]. Once a lengthy, expensive trial is complete, developers worry that the AI output might not have been sufficiently user-friendly to influence user decisions, so the results may be negative - even though a small tweak to advice wording might have made it positive.

So, in addition to algorithm validation and field trials, AI developers need to be aware of a more affordable kind of evaluation study: the lab test of algorithm impact on simulated user decisions, a kind of "intervention modelling experiment" [7]. In this kind of study, users read a clinical vignette (a scenario which may include an image or lab data) then answer questions about it (eg. What is your diagnosis?) without the AI, then re-answer the questions with the AI's output [8]. The AI output can even be presented in different formats to test the impact of alternative wording, presentation format, timing etc on user decisions. Developers do not even require a working AI algorithm for such a study, as the advice can be mocked up [8].

These user impact studies need to be designed carefully but have several benefits:

1. They are quick to set up and cheap to run
2. They help the developer estimate the likely impact of the DSS on user decisions
3. They can distinguish between effective and ineffective output formats
4. They allow the investigator to capture qualitative views on the advice [eg. 8]
5. They provide the data needed for an RCT sample size calculation.

We will discuss some examples of this kind of study [8, 9], including lessons learned from setting them up and analysing the results.

In conclusion, we believe that developers should make more use of this cost-effective study design to estimate the impact of their AI on simulated user decisions, before they make the jump to carrying out an RCT of actual impact in a real-world setting. These studies will make it much easier to develop AI algorithms with real world impact, and to identify those that are unlikely to succeed much earlier in the development pathway.

[552 words]

References

1. Global clinical AI dashboard. https://aiforhealth.app/
2. Enrico Coiera. The Last Mile: Where Artificial Intelligence Meets Reality. J *Med Internet Res 2019;21(11):e16323)* doi: 10.2196/16323
3. Petkus H, Hoogewerf J, Wyatt JC. What do senior physicians think about AI and clinical decision support systems: Quantitative and qualitative analysis of data from specialty societies. Clin Med (Lond). 2020 May;20(3):324-328. doi: 10.7861/clinmed.2019-0317.PMID: 32414724 Free PMC article.
4. Jones C, Thornton J, Wyatt JC. Enhancing trust in clinical decision support systems: a framework for developers. BMJ Health Care Inform. 2021 Jun;28(1):e100247. doi: 10.1136/bmjhci-2020-100247.PMID: 34088721 Free PMC article.
5. Boag et al. The algorithm journey map: a tangible approach to implementing AI solutions in healthcare. NPJ Digit Med. 2024;7(1):87.
6. Wyatt J, Spiegelhalter D. Field trials of medical decision-aids: potential problems and solutions. Proc Annu Symp Comput Appl Med Care. 1991:3-7.
7. Treweek S, et al. A primary care Web-based Intervention Modeling Experiment replicated behavior changes seen in earlier paper-based experiment. J Clin Epidemiol. 2016 Dec;80:116-122. doi: 10.1016/j.jclinepi.2016.07.008.
8. Scott GP, Shah P, Wyatt JC, Makubate B, Cross FW. Making electronic prescribing alerts more effective: scenario-based experimental study in junior doctors. J Am Med Inform Assoc. 2011 Nov-Dec;18(6):789-98. doi: 10.1136/amiajnl-2011-000199.
9. Dhesi AS, Wyatt JC, Estcourt LJ, McSporran W, Allard S. Insights from developing and evaluating the NHS blood choices transfusion app to support junior doctor decision making against guidelines. Transfus Med. 2022 Aug;32(4):318-326. doi: 10.1111/tme.12872.