DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence

As an increasing number of clinical decision-support systems driven by artificial intelligence progress from development to implementation, better guidance on the reporting of human factors and early-stage clinical evaluation is needed.

To the Editor—Recent years have seen an exponential growth in the number of artificial intelligence (AI) algorithms published in the medical literature, yet clinical impact in terms of patient outcomes remains to be demonstrated. One likely explanation for this so-called 'AI chasm'[1] is an overemphasis on the technical aspects of the proposed algorithms, with insufficient attention given to the factors that affect the interaction with their human users. As clinicians occupy, and are likely to keep occupying, the central role in patient care, it is essential to focus the development and evaluation of AI-based clinical algorithms on their potential to augment rather than replace human intelligence. However, AI-based decision-support systems pose unique challenges to the traditional medical decision-making process, such as their frequent lack of explainability (the so-called 'black box' problem) or their tendency to sometimes produce unexpected results. Hence, bridging algorithm development to bedside application while keeping humans at the center of the design and evaluation process is a complicated task, and current guidance is incomplete.

We make the case for a robust early and small-scale clinical evaluation stage, between the in silico algorithm development/validation (covered by the upcoming TRIPOD-AI statement[2] and STARD-AI statement[3]) and large-scale clinical trials evaluating AI interventions (covered by the CONSORT-AI statement[4]). This step can be compared to a phase 1/2 trial for drug development or (a much closer analogy, given the relationship between users' characteristics and the intervention's effectiveness) IDEAL stage 2a/2b for surgical innovation[5–7]. Four key arguments support the need for this intermediary development stage and its adequate reporting.

Human decision-making processes are complex and subject to many biases. It cannot be expected, even in the case of directive models, that human users will exactly follow all of an algorithm's recommendations, especially if these users remain accountable for their decisions[8]. In order to accurately evaluate an algorithm's performance and avoid the research waste of conducting expensive large-scale trials with decision-support systems whose interaction with human users is inadequate, it is essential to assess the actual impact of an algorithm on its users' decisions at an early stage. Additionally, consideration should be given to the difference between the development population and the target patient population, to ensure the algorithm's relevance in the implementation settings. Therefore, the assisted human performance and algorithm usability (not merely the algorithm's stand-alone outputs) need to be evaluated in the target clinical environment and need to be reported as outcomes.

Because it cannot be assumed that users' decisions will mirror the algorithm's recommendations, it is also crucially important to test the safety profile of new algorithms not only in silico but also when used to influence human decisions. Skipping this step and moving directly forward to large-scale trials would expose a considerable number of patients to an unknown risk of harm, which is ethically unacceptable. Suboptimal safety standards led to disastrous consequences in the early days of pharmacological trials; there is no need to repeat these mistakes with clinical AI.

The evaluation of human factors (ergonomics) should happen as early as possible and needs iterative evaluation–design cycles. Technical requirements often evolve as a system starts being used, and users' expectations of a system also vary in the initial exposure period. For example, users might wish for an additional key variable to make sense of the algorithm's recommendations, which in turn would require developers to access a totally different section of the electronic patient record. From an economic viewpoint, the sooner the evaluation of human factors occurs, the more cost-effective it is likely to be. Finally, iterative design modification is difficult and inappropriate during large-scale trials. Such modification would indeed cause a serious risk of invalidating the summative evaluation's conclusions, as the intervention tested is likely to have changed during trial. Early formative evaluation and rapid prototyping are therefore essential before large-scale trials.

Large-scale clinical trials are complex and expensive endeavors that require careful preparation. A well-thought-out design is essential for the production of valid and meaningful conclusions and needs background information about the intervention under evaluation. Not all such background information can be inferred from in silico evaluation, and some data have to be collected in small-scale prospective studies. For example, the most appropriate outcomes for the trial, the expected effect size, the optimal inclusion and exclusion criteria for the user population, the evolution of the users' trust in the algorithm, and the most appropriate timing of decision support are crucial pieces of information that should be known to the investigators at the time trial protocols are drafted, and these could be derived from early formative evaluation. Other important considerations, such as how to best use the output of the algorithm or how this output is to be communicated to the patients, could also be investigated at this stage.

We believe that clear and transparent reporting on these aspects will not only avoid preventable harm and research waste but also play a key role in transforming AI from a promising technology to an evidence-based component of modern medicine. This is why we have started a Delphi process[9,10] to reach expert consensus on the key information items that should be reported during 'Developmental and Exploratory Clinical Investigation of DEcision-support systems driven by Artificial Intelligence' (DECIDE-AI). The creation of the DECIDE-AI guidelines will be an open and transparent process, and we

will welcome expressions of interest from experts who wish to contribute. ❐

The DECIDE-AI Steering Group

Baptiste Vasey[1]✉, David A. Clifton[2], Gary S. Collins[3,4], Alastair K. Denniston[5], Livia Faes[6,7], Bart F. Geerts[8], Xiaoxuan Liu[5,7], Lauren Morgan[9], Peter Watkinson[10] and Peter McCulloch[1]

[1]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK. [2]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK. [3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. [4]NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK. [5]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [6]Eye Clinic, Cantonal Hospital Lucerne, Lucerne, Switzerland. [7]Moorfields Eye Hospital NHS Foundation Trust, London, UK. [8]Healthplus.ai B.V., Amsterdam, The Netherlands. [9]Morgan Human Systems, Shrewsbury, UK. [10]Critical Care Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. ✉e-mail: baptiste.vasey@nds.ox.ac.uk

Published online: 1 February 2021
https://doi.org/10.1038/s41591-021-01229-5

References

1. Keane, P. A. & Topol, E. J. *npj Digital Med.* **1**, 40 (2018).
2. Collins, G. S. & Moons, K. G. M. *Lancet* **393**, 1577–1579 (2019).
3. Sounderajah, V. et al. *Nat. Med.* **26**, 807–808 (2020).
4. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. *Br. Med. J.* **370**, m3164 (2020).
5. McCulloch, P. et al. *Lancet* **374**, 1105–1112 (2009).
6. Hirst, A. et al. *Ann. Surg.* **269**, 211–220 (2019).
7. Bilbro, N. A. et al. *Ann. Surg.* **273**, 82–85 (2021).
8. Price, W. N. II, Gerke, S. & Cohen, I. G. *J. Am. Med. Assoc.* **322**, 1765–1766 (2019).
9. Dalkey, N. & Helmer, O. *Manage. Sci.* **9**, 458–467 (1963).
10. Powell, C. *J. Adv. Nurs.* **41**, 376–382 (2003).

Competing interests

X.L. is an industry fellow (observer) with Hardian Health. D.A.C. declares academic grants from GlaxoSmithKline and personal fees from Oxford University Innovation, Biobeats and Sensyne Health, outside the context of this work. B.F.G. is CEO and Founder of Healthplus.ai and declares consulting fees from NLC Ventures Netherlands. L.M. is director and owner of Morgan Human Systems. P.W. reports grants from National Institute for Health Research, grants from Wellcome, grants from Sensyne Health, and personal fees from Sensyne Health, outside the submitted work.

Check for updates

Antibiotic resistance: a call to action to prevent the next epidemic of inequality

To the Editor — The COVID-19 pandemic has revealed the deadly impacts of structural racism and systemic health inequalities on racial and ethnic minorities in the USA. Black and Hispanic/Latinx populations have been disproportionately impacted by COVID-19, accounting for nearly half of the cases and 37% of the deaths so far, despite making up less than a third of the US population[1]. This stark imbalance has highlighted the need to examine the role of racial and ethnic disparities in shaping health outcomes.

Antibiotic resistance (AR) is widely considered to be the next global pandemic. When bacteria no longer respond to antibiotics, treatment is more costly and burdensome and is much less likely to succeed. As many as 162,000 US adults die from multidrug-resistant bacterial infections each year, which makes resistant infections the third leading cause of death[2]. Rising concerns about both the health impacts and economic impacts of AR have led to national efforts to increase surveillance, minimize inappropriate antibiotic use, jumpstart the development of diagnostics and antibiotics, and increase awareness of AR. However, the idea that AR could disproportionately impact racial and ethnic minorities has not yet entered the scientific discourse.

The existing literature describing racial and ethnic disparities in antibiotic-resistant infections in the USA is scarce and conflicting. Racial and ethnic data are not routinely collected or checked for accuracy in many clinical settings. Of the few existing studies, some suggest that Black, Hispanic and lower-income people are at higher risk of infection with community-acquired antibiotic-resistant pathogens such as methicillin-resistant *Staphylococcus aureus* and drug-resistant *Streptococcus pneumoniae*[3,4]. However, such studies are exceptionally rare. While federal efforts in the past decade have made progress in standardizing the collection and reporting of race and ethnicity data in healthcare settings, many AR-related studies still lack these data.

Nevertheless, there are a number of reasons to suspect that disparities in AR-related morbidity and mortality exist (Fig. 1). For example, while non-Hispanic Black people, Hispanic people and Asian people may receive fewer antibiotic prescriptions over their lifetimes than do non-Hispanic whites[5], they may also be more likely to consume non-prescription antibiotics[6]. Living in crowded and/or multigenerational housing, which is more common among racial and ethnic minorities[7], increases risks of AR acquisition and transmission. Some minority groups may also frequently travel to their native countries, many of which have a high burden of resistant infections[8]. Nearly 60% of people working in US meat-processing plants are Black or Hispanic/Latinx[9]; occupational contact with 'food animals' may also increase minorities' exposure to zoonotic, resistant pathogens. Finally, with more-frequent underlying comorbidities, racial and ethnic minorities are hospitalized for preventable conditions more often, which puts them at increased risk for drug-resistant hospital-acquired infections. Despite this, the US government's new National Action Plan for Combating Antibiotic-Resistant Bacteria has not prioritized racial or ethnic disparities in AR-related outcomes for either investigation or intervention[10].

As scientists, researchers and citizens, we have an obligation to ensure that racial and ethnic minorities and economically disadvantaged people will not be disproportionately burdened by the AR crisis. First, we urgently need to understand the scale of underlying disparities in AR-related morbidity and mortality. Continued improvements in the collection of racial and ethnic data in healthcare settings will enable us to evaluate factors underlying disparities across different settings and levels of 'urbanicity'. Second, we must improve AR literacy in low-income and minority communities by incorporating AR- and infection-prevention education into non-traditional settings. Tailoring future interventions to community settings such as bodegas, tiendas, daycares and classrooms, for example, could help curb unnecessary antibiotic use. Third, we must acknowledge that race or ethnicity is only one factor that might underlie disparities in AR. People who