Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group

To the Editor — Artificial intelligence (AI)-based technologies dominate medical headlines and are routinely touted as the panacea for a number of longstanding deficiencies across health systems globally. Stakeholders from healthcare, government, computer science and industry backgrounds are confident that AI can be positioned to tackle (1) the high rate of avoidable medical errors, (2) workflow inefficiencies and (3) delivery inefficiencies associated with contemporary healthcare provision[1]. Despite these lofty ambitions, the integration of AI into everyday practice within the health sector has been limited thus far.

So far, the majority of AI interventions that are close to translation are predominantly in the field of medical diagnostics. In the current paradigm, diagnostic investigations require timely interpretation from an expert clinician in order to generate a diagnosis and to direct subsequent episodes of care. The recurring issue with this system is that diagnostic services are inundated with large volumes of work, which can often supersede workforce capacity. In order to address this, diagnostic AI algorithms, with notable examples targeted toward breast cancer, lung cancer and diabetic retinopathy, have positioned themselves as medical devices[2] that may achieve diagnostic accuracy comparable to that of an expert clinician while concurrently alleviating health-resource use. However, although this paradigm shift may seem imminent, it is crucial to note that much of the evidence supporting diagnostic algorithms has been disseminated in the absence of AI-specific reporting guidelines. Without this guidance, and in a relatively nascent areas, key stakeholders are poorly placed to appraise quality and compare diagnostic accuracy between studies.

The STARD (Standards for Reporting of Diagnostic Accuracy Studies) 2015 statement[3] remains the most widely accepted set of reporting standards for diagnostic accuracy studies. In particular, STARD was developed to improve the completeness and transparency of studies investigating diagnostic accuracy. However, STARD was not designed to address the issues and challenges raised by AI-driven modalities. Such issues include unclear methodological interpretation (e.g., the use of external validation datasets, complexities of datasets and comparison to human performance) and the lack of standardized nomenclature (e.g., the definition of a 'validation dataset'), as well as the heterogeneity of outcome measures (e.g., area under the receiver operating characteristics (AUROC), sensitivity, positive predictive value and F1 score). Until these issues are overcome at a validation stage, downstream evaluation of these technologies and their potential real-word benefits will remain limited. Journal editors have also commented that approximately 25% of all manuscript submissions[4] in leading journals now center on the diagnostic accuracy of AI algorithms. In summation, there is a clear multi-faceted need to establish guidelines on the conduct and reporting of such projects.

In order to tackle these problems, the STARD-AI Steering Group is preparing an AI-specific extension to the STARD statement (STARD-AI)[5] that aims to focus upon the specific reporting of AI diagnostic accuracy studies. This work is complementary to the other novel AI extension to the EQUATOR (Enhancing Quality and Transparency of Health Research) network program, such as CONSORT-AI[6] (Consolidated Standards of Reporting Trials), SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials) and TRIPOD-ML (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis)[7]. This process is being developed in close collaboration with key stakeholders consisting of clinicians, computer scientists, journal editors, researchers, trialists, industry leaders, regulators, funders, policy makers and patient groups. In creating the STARD-AI, we view that connecting all of these groups is critical. We aim to achieve a prevailing benchmark that fulfills the recent open calls from governments, regulatory authorities, industry and academia to formulate reporting guidelines for novel AI diagnostic solutions. We are bringing together an inclusive group of representative expert stakeholders from all these areas to generate a consensus applicable across this community.

Furthermore, as a central aspect of our guideline development, we have engaged groups from typically underrepresented regions, such as Asia and Africa, in order to ensure that the AI extension to STARD will be viewed as applicable across a global scale.

We anticipate that the publication of the final recommendations will be in late 2020. In the months leading up to this, we will host a consensus meeting in which we would welcome experts from any of the aforementioned sectors to contribute their opinions on how this extension may achieve maximal coverage across specialties and backgrounds. Once published, we hope that early consultation with this STARD-AI guideline, in sync with other EQUATOR initiatives, can help direct health and life-science endeavors in this field toward the communal goal of improving the transparency and merit of AI-based evidence, as well as ultimately improving the quality of patient care. ❐

Viknesh Sounderajah[1,2], Hutan Ashrafian[1,2]✉, Ravi Aggarwal[1,2], Jeffrey De Fauw[3], Alastair K. Denniston[4,5,6], Felix Greaves[7], Alan Karthikesalingam[1,2], Dominic King[1,2], Xiaoxuan Liu[4,5,6], Sheraz R. Markar[2], Matthew D. F. McInnes[8,9], Trishan Panch[10], Jonathan Pearson-Stuttard[11], Daniel S. W. Ting[12], Robert M. Golub[13], David Moher[14], Patrick M. Bossuyt[15] and Ara Darzi[1,2]

[1]Institute of Global Health Innovation, Imperial College London, London, UK. [2]Department of Surgery and Cancer, Imperial College London, London, UK. [3]DeepMind, London, UK. [4]Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [5]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [6]Health Data Research UK, London, UK. [7]Public Health England, London, UK. [8]Department of Radiology, University of Ottawa, Ottawa, Canada. [9]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada. [10]Division of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [11]School of Public Health, Imperial College London, London, UK. [12]Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore. [13]Journal of the American Medical Association, Chicago, IL,

USA. [14]*Ottawa Hospital Research Institute, Ottawa, Canada.* [15]*Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, Amsterdam, the Netherlands.*
✉*e-mail: h.ashrafian@imperial.ac.uk*

Published online: 8 June 2020
https://doi.org/10.1038/s41591-020-0941-1

References
1. Topol, E. J. *Nat. Med.* **25**, 44–56 (2019).
2. US Food and Drug Administration. https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd (2018).
3. Bossuyt, P. M. et al. *Br. Med. J.* **351**, h5527 (2015).
4. Bluemke, D. A. et al. *Radiology* https://doi.org/10.1148/radiol.2019192515 (2019).
5. The EQUATOR Network. https://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-other-study-designs/#STARDAI (2019).
6. Liu, X. et al. *Nat. Med.* **25**, 1467–1468 (2019).
7. Collins, G. S. & Moons, K. G. M. *Lancet* **393**, 1577–1579 (2019).

Acknowledgements
Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre.

Competing interests
D.K., J.D.F. and A.K. are employees of Google Health. A.D. is an adviser at Google DeepMind/Health.

Check for updates

The QT interval in patients with COVID-19 treated with hydroxychloroquine and azithromycin

To the Editor — The SARS-CoV-2 pandemic has caused more than 1.6 million positive cases and more than 95,000 confirmed deaths as of 10 April 2020 (ref. [1]). Although there are no approved drugs to prevent or treat SARS-CoV-2 infection[2], a recent report suggested that the combination of hydroxychloroquine and azithromycin (HY/AZ) may have a favorable effect on the clinical outcomes and viral loads of infected patients[3]; this resulted in massive adoption of the regimen by clinicians worldwide. However, both medications have been independently shown to increase the risk in other populations for QT-interval prolongation, drug-induced *torsades de pointes* (a form of polymorphic ventricular tachycardia) and drug-induced sudden cardiac death[4–6]. In our center, patients with the respiratory syndrome COVID-19 who are admitted for lower airway disease with features such as non-resolving cough, chest infiltrates on X-ray and persistent fever, with or without blood-oxygen desaturation, are treated with HY/AZ. We reviewed the charts and followed the corrected QT (QTc) interval in a consecutive cohort of 84 patients receiving the regimen. HY and AZ were administered orally for 5 days. HY was given at a dose of 400 mg twice daily on the first day, followed by 200 mg twice daily. AZ was given at a dose of 500 mg per day. The average time of electrocardiograph (ECG) follow-up after HY/AZ exposure was 4.3 ± 1.7 days.

We observed prolongation of the QTc from a baseline average of 435 ± 24 ms (mean ± s.d.) to a maximal average value of 463 ± 32 ms ($P < 0.001$ (one-sample *t*-test)), which occurred on day 3.6 ± 1.6 of therapy (Fig. 1). In a subset of nine (11%) of those patients, the QTc was severely prolonged to >500 ms, a known marker of high risk of malignant arrhythmia and sudden cardiac death[7]. In this high-risk group, the QTc increased from a baseline average of 447 ± 30 ms to 527 ± 17 ms ($P < 0.01$ (one-sample *t*-test)). There were no *torsades de pointes* events recorded for any patients, including those with a severely prolonged QTc. Four patients died from multi-organ failure, without evidence of arrhythmia and without severe QTc prolongation. 64 patients remained admitted and 16 patients were discharged. The clinical and epidemiological characteristics are presented in Supplementary Table 1.

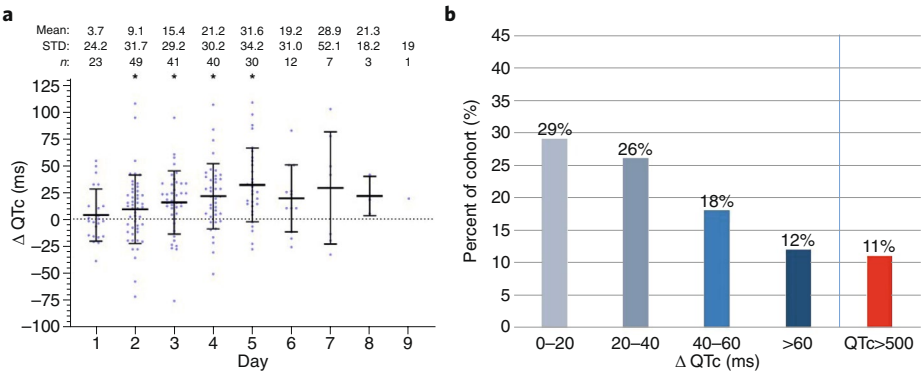

Fig. 1 | Changes in QTc on HY/AZ therapy. a, Change in QTc, presented as days after HY/AZ initiation. *$P < 0.01$, QTc compared with baseline QTc (one-sample *t*-test to compare each sample against a change in QTc (ΔQTc) of 0 ms (i.e., no change from baseline), with adjustment for multiple testing). Each data point represents a single patient with a single ECG at any given interval (*n*). STD, standard deviation. **b**, Frequency of patients with various ranges of QTc prolongation (horizontal axis). Five cardiologists trained and experienced in QT measurement performed all ECG measurements. QT and RR measurements were validated by a senior cardiac electrophysiologist expert in QT measurements. QTc was corrected with the Bazett formula ($QTc = QT/RR^{1/2}$).

The effectiveness of HZ/AZ in treating SARS-CoV-2 infection has been demonstrated in one small human study so far[2]. Previously, the combination of HY/AZ resulted in mild QTc prolongation when given to young healthy volunteers[8]. In our work, we found that in patients with COVID-19 who were treated with HY/AZ, the QTc was significantly prolonged. This discrepancy suggests that QT prolongation may be influenced by patient attributes such as the presence of co-morbidities and the severity of the disease[9]. Of note, recent guidance suggested ECG screening with QTc assessment for patients with COVID-19 who are candidates for novel therapies, including HY/AZ[10]. In our cohort, five of nine patients with severe QTc prolongation had a normal QTc at baseline. We therefore suggest that the QTc should be followed repeatedly in patients with COVID-19 who are treated with HY/AZ, particularly in those with co-morbidities and in those who are treated with other QT-prolonging medications.

Ethics declaration

The study was performed according to our Institutional Review Board guidance in