

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

Developing a Reporting Guideline for Artificial Intelligence Centred Diagnostic Accuracy Studies: The STARD-AI Protocol

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-047709
Article Type:	Protocol
Date Submitted by the Author:	06-Dec-2020
Complete List of Authors:	<p>Sounderajah, Viknesh; Imperial College London, Department of Surgery and Cancer Ashrafian, Hutan; Imperial College London, Department of Surgery and Cancer; Imperial College London, Department of Surgery and Cancer Golub, Robert; Journal of the American Medical Association Shetty, Shravya; Google Health De Fauw, Jeffrey; DeepMind Technologies Ltd Hooft, Lotty; University Medical Center Utrecht, University of Utrecht, Cochrane Netherlands Moons, Karel; Julius Center for Health Sciences and Primary Care, Epidemiology Collins, Gary; University of Oxford, Centre for Statistics in Medicine Moher, David; Ottawa Hospital Research Institute, Ottawa Methods Centre Bossuyt, Patrick M; Amsterdam University Medical Centres Darzi, Ara; Imperial College London, Institute of Global Health Innovation Karthikesalingam, Alan; Google Health Denniston, Alastair; Queen Elizabeth Hospital Birmingham, UK Mateen, Bilal Akhter; The Alan Turing Institute, Ting, Daniel; Duke-NUS Medical School, Treanor, Darren; University of Leeds King, Dominic; Imperial College London, Centre for Health Policy Greaves, Felix; Imperial College London, Department of Primary Care and Public Health Godwin, Jonathan; DeepMind Technologies Ltd Pearson-Stuttard, Jonathan; Imperial College London, PCPH Harling, Leanne; Imperial College London, Department of Surgery and Cancer McInnes, Matthew; University of Ottawa, Rifai, Nader; Harvard Medical School, Tomasev, Nenad; DeepMind Technologies Ltd Normahani, Pasha; Imperial College London, Department of Surgery and Cancer Aggarwal, Ravi; Imperial College London, Department of Surgery and Cancer Markar, Sheraz; Imperial College London, Vollmer, Sebastian; The Alan Turing Institute Panch, Trishan</p>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

	Liu, Xiaoxuan; University of Birmingham Whiting, Penny; University of Bristol, School of Social and Community Medicine
Keywords:	Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT



BMJ Open: first published as 10.1136/bmjopen-2020-047709 on 28 June 2021. Downloaded from <http://bmjopen.bmj.com/> on June 10, 2023 by guest. Protected by copyright.



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Developing a Reporting Guideline for Artificial Intelligence Centred Diagnostic Accuracy Studies:

The STARD-AI Protocol

Authors:

Viknesh Sounderajah^{1,2}, Hutan Ashrafian^{1,2}, Robert Golub¹¹, Shravya Shetty⁶, Jeffrey De Fauw³, Lotty Hooft¹⁸, Carl Moons¹⁸, Gary Collins¹⁷, David Moher¹², Patrick Bossuyt¹³ and Ara Darzi^{1,2} on behalf of the STARD-AI Steering Committee (Alan Karthikesalingam⁶, Alastair Denniston^{4,15,16}, Bilal Mateen¹⁸, Daniel Ting¹⁰, Darren Treanor²⁰, Dominic King²¹, Felix Greaves⁵, Jonathan Godwin³, Jonathan Pearson-Stuttard⁹, Leanne Harling², Matthew McInnes⁷, Nader Rifai²², Nenad Tomasev³, Pasha Normahani², Penny Whiting²³, Ravi Aggarwal¹, Sebastian Vollmer¹⁹, Sheraz Markar², Trishan Panch⁸ and Xiaoxuan Liu^{4,15,16})

Author Affiliations

¹ Institute of Global Health Innovation, Imperial College London, United Kingdom

² Department of Surgery and Cancer, Imperial College London, United Kingdom

³ DeepMind, United Kingdom

⁴ Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, United Kingdom

⁵ The National Institute for Health and Care Excellence, United Kingdom

⁶ Google Health

⁷ Department of Radiology, University of Ottawa, Canada

⁸ Division of Health Policy and Management, Harvard T.H. Chan School of Public Health, United States of America

⁹ School of Public Health, Imperial College London, United Kingdom

¹⁰ Singapore Eye Research Institute, Singapore National Eye Center, Singapore

¹¹ JAMA (Journal of the American Medical Association), United States of America

1
2
3 ¹² Ottawa Hospital Research Institute, Canada
4

5
6 ¹³ Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam,
7
8 The Netherlands
9

10 ¹⁴ University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom
11

12 ¹⁵ Health Data Research UK, London, United Kingdom
13

14 ¹⁶ Clinical Epidemiology Program, Ottawa Hospital Research Institute, Canada
15

16
17 ¹⁷ Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and
18
19 Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, United Kingdom
20

21 ¹⁸ Julius Center for Health Sciences and Primary Care, and Cochrane Netherlands, University Medical
22
23 Center Utrecht, Utrecht University, Utrecht, The Netherlands
24

25
26 ¹⁹ Alan Turing Institute, Kings Cross, United Kingdom
27

28 ²⁰ Leeds Teaching Hospitals NHS Trust, University of Leeds, Leeds, United Kingdom
29

30 ²¹ Optum, Paddington, London, United Kingdom
31

32 ²² Department of Laboratory Medicine, Boston Children's Hospital, Harvard Medical School, Boston,
33
34 Massachusetts, United States of America
35

36
37 ²³ School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom
38
39

40 **Author disclosures:**

41
42 The views and opinions expressed herein are those of the authors and do not necessarily reflect the
43
44 views of their employers or funders.
45
46
47
48

49 **Corresponding author:**

50
51 Mr Hutan Ashrafian BSc (Hons) MBBS MRCS PhD MBA
52

53
54 Institute of Global Health Innovation, 10th Floor, Queen Elizabeth Queen Mother building, St Mary's
55
56 Hospital Campus, Praed Street, London, United Kingdom, W2 1NY
57

58 **Telephone Number:** +447799871597
59
60

1
2
3 **E-mail:** hutan@imperial.ac.uk
4
5
6

7
8 **Funding:**
9

10 Infrastructure support for this research was provided by the NIHR Imperial Biomedical Research
11 Centre (BRC).

12
13
14 GSC is supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK
15 (programme grant: C49297/A27294).
16

17
18
19 DT is funded by National Pathology Imaging Co-operative, NPIC (Project no. 104687) is supported by
20 a £50m investment from the Data to Early Diagnosis and Precision Medicine strand of the
21 government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and
22 Innovation (UKRI).
23
24
25
26

27
28 FG is supported by the National Institute for Health Research Applied Research Collaboration
29 Northwest London
30
31

32
33
34 **Data Statement:**
35

36 There is no data in this work.
37
38
39

40
41 **Word count (main body):**
42

43 3201
44
45
46
47

48 **Study Status:**
49

50 Stage 2 of this study has been completed. Stage 3 (the modified Delphi consensus process) is
51 underway.
52
53
54
55
56
57
58
59
60

Abstract

Introduction:

STARD was developed to improve the completeness and transparency of reporting in studies investigating diagnostic accuracy. However, its current form, STARD 2015 does not address the unique issues and challenges raised by artificial intelligence (AI) centred interventions. As such, we propose an AI-specific version of the STARD checklist (STARD-AI 2021), which focuses upon the reporting of AI diagnostic accuracy studies. This paper describes the processes and methods that will be used to develop STARD-AI.

Methods and analysis:

Following guidance from the EQUATOR network, the development of the STARD-AI 2021 checklist can be distilled into six stages. (1) A project organisation phase has been undertaken, during which a Project Team and a Steering Committee were established. (2) An item generation process has been completed following a literature review, a patient and public involvement and engagement (PPIE) exercise and an online scoping survey of international experts. (3) A three-round modified Delphi consensus methodology is proposed, which will culminate in a teleconference consensus meeting of experts. (4) Thereafter, the Project Team will draft the initial STARD-AI checklist and the accompanying statement. (5) A piloting phase amongst expert and non-expert users will be carried out to identify items which are considered to be unclear, ambiguous or missing. This process, consisting of surveys and interviews, will contribute towards the explanation and elaboration document. (6) Upon finalisation of the manuscripts, a further teleconference meeting between the Project Team and Steering Committee is proposed prior to dissemination and implementation.

Ethics and dissemination:

Ethical approval has been granted by the Joint Research Compliance Office at Imperial College London (SETREC reference number: 19IC5679). A tailored dissemination strategy will be aimed towards 5

1
2
3 groups of stakeholders: (a) academia, (b) policy, (c) guidelines and regulation, (d) industry and (e)
4
5 public and non-specific stakeholders. We anticipate that dissemination will take place in Q2 of 2021.
6
7
8
9

10 Key words:

11
12 Diagnostic accuracy, reporting guideline, artificial intelligence, STARD, transparency
13
14
15
16

17 Word count: 300/300
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Article Summary

Strengths and limitations of this study:

- Gap: There are no specific reporting standards for artificial intelligence (AI) diagnostic accuracy studies
- Solution: We are developing a specific set of reporting standards for AI diagnostic accuracy studies; STARD-AI 2021.
- Clinical implications: This will help key stakeholders to appraise quality and compare diagnostic accuracy of AI models that are reported scientific studies.
- Strengths: STARD-AI 2021 will be a product of extensive evidence generation process that is led by multiple stakeholders (clinician scientists, computer scientists, journal editors, EQUATOR Network representatives, reporting guideline developers, epidemiologists, statisticians, industry leaders, funders, health policy makers, patients, legal experts and medical ethicists).
- Limitations: views of Delphi panellists may differ from those experts who decline participation.

Only

Glossary

Project Team

This consists of the founder of STARD (PMB), the former United Kingdom Minister for Health and the current chair for the National Health Service Accelerated Access Collaborative (AD), members of the TRIPOD-AI group (GSC, LH, KGM), a senior software engineer (SS), directors of the EQUATOR Network (DM, GSC), the scientific content deputy editor for JAMA (RG) as well as 2 clinician scientists from Imperial College London (HA (supervisor), VS (doctoral research fellow)).

Steering Committee

This consists of clinician scientists, computer scientists, journal editors, EQUATOR Network representatives, epidemiologists, statisticians, industry leaders, funders, health policy makers, legal experts and medical ethicists. These individuals were identified through their notable work with respect to (1) diagnostic accuracy research, (2) artificial intelligence in healthcare, (3) health policy, (4) contribution to AI-centred EQUATOR initiatives, such as TRIPOD-AI, CONSORT-AI and SPIRIT-AI.

Consensus Group

This consists of experts who participated in the modified Delphi consensus process of the study.

Pilot Group

This consists of experts who participated in the pilot phase (Stage 5) of the study.

Checklist

A document listing the minimally essential items that should be reported in all diagnostic accuracy studies centred around artificial intelligence interventions. This constitutes the core of the reporting guideline.

Statement

Provides the rationale in the development of this reporting guideline, describes the process of developing the checklist, the checklist, dissemination and implementation plans, and any evaluation plans.

Explanation and Elaboration (E&E)

Provides the rationale behind each item in the checklist, along with examples of good reporting.

Reporting guideline

The combination of the checklist, statement and E&E material.

Flow diagram

A flow diagram depicts the flow of information through the different phases of a study.

Artificial Intelligence (AI)

The science of developing computer systems which can perform tasks normally requiring human intelligence.

Delphi study

A research method that derives the collective opinions of a group through a staged consultation of surveys, questionnaires, or interviews, with an aim to reach consensus at the end.

Introduction

Artificial intelligence (AI) is commonly cited as an imminent disruptive innovation[1] within the health sector. If used successfully, AI has the potential to tackle (1) the high rate of avoidable medical errors, (2) workflow inefficiencies and (3) delivery inefficiencies associated with modern healthcare provision[2]. The majority of AI interventions that are close to translation are in the field of medical diagnostics[3]. In the current paradigm, diagnostic investigations require timely interpretation from an expert clinician in order to generate a diagnosis and to subsequently direct episodes of care. However, the recurring issue with the present system is that diagnostic services are inundated with large volumes of work, which often exceeds workforce capacity[4]; COVID-19 being an immediate case in point. In order to address this, diagnostic AI algorithms have positioned themselves as medical devices that may achieve diagnostic accuracy comparable to that of an expert clinician whilst concurrently alleviating health-resource use. 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 area, key stakeholders are poorly placed to appraise quality and compare diagnostic accuracy between scientific studies.

The STARD (Standards for Reporting of Diagnostic Accuracy Studies) 2015 statement remains the most widely accepted set of reporting standards for diagnostic accuracy studies[5]. STARD was developed to improve the completeness and transparency of studies investigating diagnostic accuracy. It consists of a checklist of 30 items that authors are strongly encouraged to address when reporting their diagnostic accuracy studies. It is endorsed by over 200 biomedical journals[6] and studies have shown that adherence to the STARD checklist leads to improved reporting of key study parameters[7,8].

However, in its current iteration, STARD 2015 is not designed to address the issues and challenges raised by AI-driven modalities. Issues include unclear methodological interpretation (e.g., the use of

1
2
3 external validation datasets, complexities of datasets and comparison to human performance), the
4
5 lack of standardized nomenclature (e.g., the definition of a 'validation dataset'), as well as the
6
7 heterogeneity of outcome measures (e.g., area under the receiver operating characteristics (AUROC),
8
9 sensitivity, positive predictive value and F1 score). Until these issues are overcome, achieving
10
11 comprehensive evaluations of these technologies and their potential translational benefits will remain
12
13 limited.
14
15

16
17
18 In order to tackle these problems, we propose an AI-specific STARD guideline (STARD-AI) that aims to
19
20 focus upon the reporting of AI diagnostic accuracy studies[9]. This work is complementary to the other
21
22 AI centred checklists listed in the EQUATOR (Enhancing Quality and Transparency of Health Research)
23
24 Network program (www.equator-network.org)[10], such as SPIRIT-AI (Standard Protocol Items:
25
26 Recommendations for Interventional Trials)[11], CONSORT-AI (Consolidated Standards of Reporting
27
28 Trials)[12] and TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual
29
30 Prognosis or Diagnosis)[13].
31
32

33
34
35 STARD-AI is being coordinated by a global Project Team and Steering Committee consisting of clinician
36
37 scientists, computer scientists, journal editors, EQUATOR Network representatives, reporting
38
39 guideline developers, epidemiologists, statisticians, industry leaders, funders, health policy makers,
40
41 legal experts and medical ethicists. In devising STARD-AI, we view that connecting all of these key
42
43 stakeholders across the world is of the utmost importance.
44
45
46
47

48 Aim

49
50
51
52 This study aims to produce a novel AI centred diagnostic accuracy checklist (STARD-AI) which
53
54 appropriately accounts for the specific considerations warranted in the reporting of AI diagnostic
55
56 accuracy studies.
57
58
59
60

Focus of STARD-AI

The scope of STARD-AI 2021 is to address studies that use AI techniques to assess diagnostic accuracy (or clinical performance). Such studies compare test results between individuals (typically patients) with and without a target condition (or disease). Samples or images from study participants undergo assessment by a diagnostic technique which is designed to pick-up the target condition. This occurs alongside a concomitant reference standard or “gold-standard” test for the target condition in a defined timeframe. The diagnostic technique can account for either single or combined tests and typically includes (1) imaging data (e.g. CT scans), (2) pathological data (digitised specimen slide) or (3) reporting data (e.g. electronic health records or multi-omic spectra). STARD-AI 2021 also accounts for image segmentation and data delineation between a target condition and its absence (such as normal anatomy or health record results).

Estimates of clinical performance, or accuracy, are based on a comparison of the classification based on the test results with the classification by the reference standard, or gold standard, of the same patients. Alternatively, the reference standard can be the occurrence of an event within a defined timeframe.

STARD-AI was developed to guide the reporting of evaluations of the accuracy, or performance, of AI applications. If the emphasis of the study is on developing, validating, or updating a multivariable prediction model, the TRIPOD-AI reporting guidelines (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) may be more appropriate.

Methods

This protocol has been constructed in accordance with the EQUATOR Network (Enhancing the Quality and Transparency of Health Research) toolkit for developing reporting guidelines[14]. It has also greatly benefitted from the experience and expertise from Project Team and Steering Committee members who had previously led the STARD 2003[15], STARD 2015, STARD for Abstracts[16], SPIRIT-AI and CONSORT-AI initiatives respectively.

We are able to distil the development of the STARD-AI 2021 checklist into six stages. The overall goal of the STARD-AI initiative is to generate a list of minimally essential items, based upon the established STARD 2015 framework, that should be reported in all AI diagnostic accuracy studies. The items must assist the reader to appraise the completeness, applicability and potential for bias of the study findings.

Stage 1: Project organisation

A ten member STARD-AI Project Team was established in order to coordinate the guideline development process. The Project Team consists of the founder of STARD (PMB), the former United Kingdom Minister for Health and the current chair for the National Health Service Accelerated Access Collaborative (AD), members of the TRIPOD-AI core committee (GSC, LH, KGM), a senior software engineer (SS), directors of the EQUATOR Network (DM, GSC), the scientific content deputy editor for JAMA (RG) as well as 2 clinician scientists from Imperial College London (HA (supervisor), VS (doctoral research fellow)). The Project Team are responsible for identifying suitable members of the Steering Committee, candidate item generation, undertaking the online surveys for the modified Delphi consensus process, organising the consensus meeting, drafting the STARD-AI 2021 checklist and accompanying documents, coordinating the piloting the draft STARD-AI checklist as well as leading the dissemination process.

1
2
3 Further to the Project Team, a multidisciplinary STARD-AI Steering Committee was established in order
4
5 to provide specialist guidance throughout the STARD-AI process. This committee consists of clinician
6
7 scientists, computer scientists, journal editors, EQUATOR network directors, epidemiologists,
8
9 statisticians, industry leaders, funders, health policy leaders, regulatory leaders, legal experts, patient
10
11 representation experts and medical ethicists. These individuals were identified through their notable
12
13 work with respect to (1) diagnostic accuracy research and its associated clinical translation, (2) applied
14
15 artificial intelligence in healthcare as well as (3) notable contribution to other AI-centred EQUATOR
16
17 Network registered initiatives, such as TRIPOD-AI, CONSORT-AI and SPIRIT-AI.
18
19
20
21
22

23 Prior to Stage 2, the STARD-AI project was registered with the EQUATOR Network.
24
25
26

27 Stage 2: Item generation 28 29

30
31 In order to generate a candidate list of items to enter the modified Delphi consensus process, the
32
33 Project Team undertook a literature review, an extensive online scoping survey with an international
34
35 panel of experts and a patient public involvement and engagement (PPIE) exercise.
36
37
38

39 a) Literature review: 40 41 42 43

44 In January 2020, a literature review of both academic and non-academic literature was undertaken.
45
46 An electronic database search of Medical Literature Analysis and Retrieval System Online (MEDLINE)
47
48 and Excerpta Medica database (EMBASE) was conducted through Ovid. Both Medical Subject Headings
49
50 (MeSH) or EMBASE Subject Headings (Emtree) were used. Search results will be imported into
51
52 Covidence (Covidence.org, Melbourne, Australia) for duplicate removal and study selection. Two
53
54 individuals (VS/HA) individually screened study titles and abstracts for inclusion. Disagreements were
55
56 resolved through discussion.
57
58
59
60

1
2
3 This process was augmented by non-systematic searches using traditional search engines for grey
4 literature, social networking platforms as well as personal article collections highlighted by members
5 of the Project Team. Titles and abstracts of shortlisted publications were screened by one of two
6 reviewers (VS, HA) and potentially eligible publications were retrieved for full-text assessment.
7
8 Extracted material were broadly classified into four categories by VS and HA; (1) general
9 considerations regarding diagnostic accuracy studies and artificial intelligence, (2) evidence and
10 statements suggesting modification to the STARD 2015 checklist, (3) evidence and statements
11 suggesting additions to the STARD 2015 checklist and (4) evidence and statements suggesting the
12 removal of specific items from the STARD 2015 checklist.
13
14
15
16
17
18
19
20
21
22
23
24

25 b) Online scoping survey:

26
27
28
29 In addition to this, in February 2020, the Project Team undertook an online survey with an
30 international panel of experts (n=80) in order to identify potential further items or modifications that
31 warrant consideration. This process generated over 2500 responses, which were analysed and classed
32 into the aforementioned 4 broad categories.
33
34
35
36
37
38
39

40 c) Patient public involvement and engagement (PPIE) exercise:

41
42
43
44
45 Lastly, a focus group was conducted with patients and members of the public who had expressed an
46 interest in participating in forums related to digital health and AI. The objective of these discussions
47 was two-fold; (1) to further identify issues not uncovered during the literature review and expert
48 survey and (2) to gain further understanding of the perceived importance of specific items raised thus
49 far. These discussions were conducted remotely using Zoom (Zoom Video Communications, Inc., USA).
50
51
52
53
54
55
56
57
58
59
60

1
2
3 An expert facilitator led a discussion on the current use of AI in healthcare, on what the aims of STARD-
4 AI were and what participants considered to be important items to capture during the study process.
5
6 As stakeholder discussions were conducted virtually on Zoom, anonymised post-hoc discussion
7 transcripts were maintained. Two investigators (VS, HA) independently identified common themes
8 and sub-themes from the discussion, which were classed into the aforementioned 4 broad categories.
9
10
11
12
13
14
15

16 Having synthesised the findings of the literature review, the survey and the patient public involvement
17 and engagement exercise, the Project Team, in collaboration with the Steering Committee, decided
18 upon which items warrant consideration in the formal modified Delphi consensus process.
19
20
21
22
23
24

25 Stage 3: Modified Delphi consensus process

26 a) Study design and participants:

27
28
29 We will adopt a pragmatic modified Delphi consensus methodology. The Delphi consensus
30 methodology is a well-established method[17] of obtaining a collective opinion from a group of
31 experts through a series of questionnaires; each one refined based upon feedback from respondents
32 on a previous version.
33
34
35
36
37
38
39
40
41
42
43
44

45 Participants are invited to join the STARD-AI Consensus Group on account of their expertise as clinician
46 scientists, computer scientists, journal editors, EQUATOR Network representatives, reporting
47 guideline developers, epidemiologists, statisticians, industry leaders (e.g., clinician scientists,
48 computer scientists and product managers from health technological companies), funders, health
49 policy makers, legal experts and medical ethicists. Invited experts will be provided with three weeks
50 to respond to the initial invitation to participate. Those who accept the invitation will be invited to
51 complete each round of the modified Delphi consensus process. Those who contribute to both online
52
53
54
55
56
57
58
59
60

1
2
3 rounds will be acknowledged by name as an author, within a group authorship model, in the
4
5 publication that arises from this study.
6
7
8
9

10 In each round of the modified Delphi consensus process, participants will be asked to grade each
11
12 candidate item using a 5-point Likert-like scale (1 – very important, 2 – important, 3 – moderately
13
14 important, 4 – slightly important, 5 – not at all important). The threshold for consensus will be
15
16 predefined at $\geq 80\%$. Items which achieve $\geq 80\%$ ratings of 1 or 2 will be deemed to be essential for
17
18 inclusion and will be put forward for discussion in the final round (round 3, which will occur in the
19
20 form of a virtual teleconference meeting). Items which achieve $\geq 80\%$ ratings of 4 or 5 will be deemed
21
22 unimportant for inclusion and will be excluded. Items which did not reach this threshold of consensus
23
24 will be put forward to the next round of the modified Delphi consensus process. In addition to rating
25
26 items, participants will again be asked in a free-text format to suggest any other items that they
27
28 consider to be potentially important to discuss in subsequent rounds.
29
30
31
32
33
34

35 In round 2, the survey will compose of items for which consensus was not achieved and any new items
36
37 suggested in round 1. Next to each item, participants will be reminded of what rating they gave in the
38
39 previous round. Additionally, the mean score given by the overall group in the previous round will be
40
41 displayed for each item. Thus, participants will be able to revise their initial score with the additional
42
43 knowledge of other participant responses. Following collection of round 2 responses, additional
44
45 consensus items will be put forward for discussion during round 3 whilst negative consensus items will
46
47 be excluded.
48
49
50
51

52 Any resulting non-consensus items from round 3 will again be put forward for voting in a final round,
53
54 which will occur alongside the teleconference consensus meeting. Any final non-consensus items will
55
56 then be resolved through discussion amongst those in virtual attendance at the consensus meeting.
57
58
59
60

1
2
3 b) Round 3; the consensus meeting:
4
5
6

7 The consensus meeting (round 3) will consist of the STARD-AI Project Team and the STARD-AI Steering
8 Committee. Given COVID-19 constraints, the meeting will be conducted virtually using Zoom (San Jose,
9 United States of America). The primary objective is to develop a consensual draft version of STARD-AI
10 checklist. As recommended in the COMET handbook, the nominal group technique, a highly-
11 structured group interaction framework, will be utilised to aid this process[18,19]. Following a brief
12 introduction and explanation of the purpose of the meeting by the facilitators (VS and HA),
13 participants will discuss the inclusion and exclusion of candidate items. Participants will be asked to
14 share any comments they have generated in a 'round robin' format until all contributions are
15 exhausted. Participants will then be invited to discuss or seek further clarification about any of the
16 ideas or comments produced. This discussion phase will be led by the facilitator (VS and HA) to ensure
17 that the discussion will not be dominated by any one individual and be as neutral as possible[20].
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

34 c) Study conduct:
35
36
37
38

39 VS and HA will be the Delphi facilitators for the online rounds as well as the teleconference consensus
40 meeting. They are responsible for the creation of the questionnaires, the invitations, the responses,
41 the reminders, the analysis as well as the feedback for subsequent rounds.
42
43
44
45
46
47

48 The first two rounds of the modified Delphi consensus process will be conducted as online surveys
49 using the DelphiManager software (version 4.0), which is developed and maintained by the COMET
50 (Core Outcome Measures in Effectiveness Trials) initiative. Round 3 (the consensus meeting) will be
51 carried using Zoom.
52
53
54
55
56
57
58
59
60

1
2
3 Stage 4: Development of the (1) checklist, (2) statement and (3) explanation and elaboration (E&E)
4
5 document
6
7
8

9 Upon completion of the modified Delphi consensus process, the Project Team will draft the initial
10 STARD-AI checklist and statement. The draft checklist and statement will be shared amongst the wider
11 Steering Committee in order to discuss its content and therefore allowing the Steering Committee to
12 suggest additions, subtractions or modifications as they see fit. This stage will also allow for
13 harmonisation of key terms with the imminent TRIPOD-AI, in addition to the existing CONSORT-AI and
14 SPIRIT-AI checklists.
15
16
17
18
19
20
21
22
23

24 Stage 5: Piloting amongst experts and non-experts
25
26
27

28 Upon completion of the first draft of the STARD-AI checklist, we intend to organise multiple rounds of
29 piloting amongst expert and non-expert users (Pilot Group). The main aim of these piloting sessions is
30 to identify items which are considered to be vague, ambiguous or perceived to be missing. We intend
31 to undertake this process amongst radiology experts, pathology experts, computer scientists, expert
32 statisticians, journal editorial boards, members of the global EQUATOR Network, key industry
33 stakeholders as well as policy experts. Interviews amongst this Pilot Group will be undertaken in order
34 to ensure that a granular level of feedback is attained for points of discussion. Experts and non-experts
35 within the Pilot Group will be acknowledged by name as an author, within a group authorship model,
36 in the publications that arise from this study.
37
38
39
40
41
42
43
44
45
46
47
48
49

50 In conjunction to this piloting process, the Project Team will also prepare the explanation and
51 elaboration (E&E) document, to provide rationale for the included items along with examples of good
52 reporting.
53
54
55
56
57

58 Stage 6: Finalisation, publication and post-publication activities
59
60

1
2
3 Following the piloting phase, the final proposed amendments to STARD-AI will be discussed amongst
4 the Project Team and the Steering Committee. Once consensus has been reached through e-mail
5 correspondence, the documents will be disseminated.
6
7
8
9

10
11 At this stage, a further discussion regarding the final strategy for dissemination and implementation
12 of STARD-AI will occur amongst the Project Team and the Steering Committee. We strongly anticipate
13 that the dissemination strategy will be principally tailored towards 5 groups of stakeholders; (a)
14 academia, (b) policy, (c) guidelines and regulation, (d) industry and (e) patient representing bodies.
15
16 Although a significant amount of material will cross over between stakeholders, creating stakeholder
17 specific material is considered to be the most meaningful way of achieving impact.
18
19
20
21
22
23
24

25
26
27 a) Academic stakeholders:
28
29

30
31 We aim to publish the STARD-AI checklist, the accompanying statement and the E&E document in an
32 open access format in a high-impact peer-reviewed journal. We will also share all relevant material
33 through the EQUATOR website. In order to further complement this, we aim to create specialty-
34 specific discourse regarding STARD-AI through focussed editorials in pertinent journals. These journal
35 editors will also be actively encouraged to endorse STARD-AI as part of their broader editorial policy.
36
37 Moreover, we will present STARD-AI at national and international scientific meetings. Translations of
38 the guideline in various languages are actively encouraged in order to further broaden the scope of its
39 impact. We encourage interested parties to contact the corresponding author for further information
40 about the translation policies.
41
42
43
44
45
46
47
48
49
50

51
52
53 b) Policy stakeholders:
54
55

56
57 We aim to persuade governmental bodies to adopt the checklist as part of their policy assessments.
58
59 This will involve presentations at national and international health policy summits (e.g., World
60

1
2
3 Innovation Summit for Health, NHS Accelerated Access Collaborative, National Institutes of Health).
4
5 Furthermore, we will aim to integrate teaching about STARD-AI into national health policy educational
6
7 programmes (the master's programme (MSc) for Health Policy at Imperial College London, the NHS
8
9 Digital Academy, UK Research Innovation Centres of Excellence in AI in Digital Imaging).
10
11
12
13

14 c) Guidelines and regulatory stakeholders:
15
16

17
18 We aim to work alongside guidelines and regulatory bodies to adopt the checklist as part of their
19
20 national health technology assessments. This will involve the United States Food and Drug
21
22 Administration (FDA), the Medicines and Healthcare products Regulatory Agency (MHRA), The
23
24 National Institute for Health and Care Excellence (NICE), the Horizon 2020 programme, the European
25
26 Medicines Agency as well as the Consortia for Improving Medicine with Innovation and Technology
27
28 (CIMIT).
29
30

31
32
33 d) Industry stakeholders:
34
35

36
37 We will present STARD-AI to a broad range of health technology companies (ranging from start-ups,
38
39 small and medium-sized enterprises to multinational corporations) so that their product pipelines may
40
41 accommodate for this.
42
43

44
45 e) Public and non-specific stakeholders:
46
47

48
49 Ensuring that the core material (STARD-AI checklist, statement and explanation and elaboration
50
51 document) is available in an open access fashion, through a CC-BY license, is paramount to achieving
52
53 general impact. In addition, we aim to publish articles in mainstream media and attain distribution
54
55 through non-traditional means (e.g. social networking platforms, webinars, podcast episodes and blog
56
57 posts).
58
59
60

Ethics

Ethical approval has been granted by the Joint Research Compliance Office at Imperial College London (SETREC reference number: 19IC5679).

Author Statement

Viknesh Sounderajah, Hutan Ashrafian, Robert Golub, Shravya Shetty, Jeffrey De Fauw, Lotty Hooft, Carl Moons, Gary Collins, David Moher, Patrick Bossuyt and Ara Darzi were involved in the planning and design of the study. Viknesh drafted the manuscript with all authors contributing to the writing.

Alan Karthikesalingam, Alastair Denniston, Bilal Mateen, Daniel Ting, Darren Treanor, Dominic King, Felix Greaves, Jonathan Godwin, Jonathan Pearson-Stuttard, Leanne Harling, Matthew McInnes, Nader Rifai, Nenad Tomasev, Pasha Normahani, Penny Whiting, Ravi Aggarwal, Sebastian Vollmer, Sheraz Markar, Trishan Panch and Xiaoxuan Liu are members of the STARD-AI Steering Committee. They are equally involved in the wider conduct and direction of the overall study. All of the authors edited the manuscript and provided critical appraisal.

All named authors approved the final draft of the manuscript.

References

- 1 Sounderajah V, Patel V, Varatharajan L, *et al*. Are disruptive innovations recognised in the healthcare literature? A systematic review. *BMJ Innov* 2020;:bmjinnov-2020-000424. doi:10.1136/bmjinnov-2020-000424
- 2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 2019;**25**:44–56. doi:10.1038/s41591-018-0300-7
- 3 Benjamens S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digit Med* 2020;**3**:118. doi:10.1038/s41746-020-00324-0
- 4 Williams BJ, Bottoms D, Treanor D. Future-proofing pathology: The case for clinical adoption of digital pathology. *J Clin Pathol* 2017;**70**:1010–8. doi:10.1136/jclinpath-2017-204644
- 5 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;**351**. doi:10.1136/bmj.h5527
- 6 Ochodo EA, Bossuyt PM. Reporting the Accuracy of Diagnostic Tests: The STARD Initiative 10 Years On. *Clin Chem* 2013;**59**:917–9. doi:10.1373/clinchem.2013.206516
- 7 Korevaar DA, Van Enst WA, Spijker R, *et al*. Reporting quality of diagnostic accuracy studies: A systematic review and meta-analysis of investigations on adherence to STARD. *Evid. Based. Med.* 2014;**19**:47–54. doi:10.1136/eb-2013-101637
- 8 Korevaar DA, Wang J, Van Enst WA, *et al*. Reporting diagnostic accuracy studies: Some improvements after 10 years of STARD. *Radiology* 2015;**274**:781–9. doi:10.1148/radiol.14141160

- 1
2
3 9 Sounderajah V, Ashrafian H, Aggarwal R, *et al*. Developing specific reporting guidelines for
4 diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat*.
5
6
7
8
9
10
11
12 10 The EQUATOR Network | Enhancing the QUALity and Transparency Of Health Research.
13
14 <https://www.equator-network.org/> (accessed 26 Sep 2020).
15
16
17
18 11 Rivera SC, Liu X, Chan A-W, *et al*. Consensus statement Guidelines for clinical trial protocols
19 for interventions involving artificial intelligence: the SPIRIT-AI extension The SPIRIT-AI and
20 CONSORT-AI Working Group*, SPIRIT-AI and CONSORT-AI Steering Group and SPIRIT-AI and
21 CONSORT-AI Consensus Group. *Nat Med* 2020;**26**:1351–63. doi:10.1038/s41591-020-1037-7
22
23
24
25
26
27
28
29 12 Liu X, Rivera SC. Consensus statement Reporting guidelines for clinical trial reports for
30 interventions involving artificial intelligence: the CONSORT-AI extension^{6,13} and The
31 SPIRIT-AI and CONSORT-AI Working Group*. *Nat Med* 2020 269 2020;**26**:1364–74.
32
33
34
35
36
37
38
39
40 13 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*.
41
42 2019;**393**:1577–9. doi:10.1016/S0140-6736(19)30037-6
43
44
45
46 14 Toolkits | The EQUATOR Network. <https://www.equator-network.org/toolkits/> (accessed 26
47
48 Sep 2020).
49
50
51
52 15 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. The STARD statement for reporting studies of
53 diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;**138**.
54
55
56
57
58
59
60

- 1
2
3 16 Cohen JF, Korevaar DA, Gatsonis CA, *et al.* STARD for Abstracts: Essential items for reporting
4 diagnostic accuracy studies in journal or conference abstracts. *BMJ* 2017;**358**.
5
6 doi:10.1136/bmj.j3751
7
8
9
10
11 17 Brown BB. Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts.
12 Published Online First: 1968.<https://www.rand.org/pubs/papers/P3925.html> (accessed 26
13 Sep 2020).
14
15
16
17
18
19 18 McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. *Int J*
20 *Clin Pharm* 2016;**38**:655–62. doi:10.1007/s11096-016-0257-x
21
22
23
24
25
26 19 Williamson PR, Altman DG, Bagley H, *et al.* The COMET Handbook: version 1.0. *Trials*
27 2017;**18**:280. doi:10.1186/s13063-017-1978-4
28
29
30
31
32 20 Harvey N, Holmes CA. Nominal group technique: An effective method for obtaining group
33 consensus. *Int J Nurs Pract* 2012;**18**:188–94. doi:10.1111/j.1440-172X.2012.02017.x
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

BMJ Open

Developing a Reporting Guideline for Artificial Intelligence Centred Diagnostic Test Accuracy Studies: The STARD-AI Protocol

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-047709.R1
Article Type:	Protocol
Date Submitted by the Author:	04-May-2021
Complete List of Authors:	<p>Sounderajah, Viknesh; Imperial College London, Department of Surgery and Cancer Ashrafian, Hutan; Imperial College London, Department of Surgery and Cancer; Imperial College London, Department of Surgery and Cancer Golub, Robert; Journal of the American Medical Association Shetty, Shravya; Google Health De Fauw, Jeffrey; DeepMind Technologies Ltd Hooft, Lotty; University Medical Center Utrecht, University of Utrecht, Cochrane Netherlands Moons, Karel; Julius Center for Health Sciences and Primary Care, Epidemiology Collins, Gary; University of Oxford, Centre for Statistics in Medicine Moher, David; Ottawa Hospital Research Institute, Ottawa Methods Centre Bossuyt, Patrick M; Amsterdam University Medical Centres Darzi, Ara; Imperial College London, Institute of Global Health Innovation Karthikesalingam, Alan; Google Health Denniston, Alastair; Queen Elizabeth Hospital Birmingham, UK Mateen, Bilal Akhter; The Alan Turing Institute, Ting, Daniel; Duke-NUS Medical School, Treanor, Darren; University of Leeds King, Dominic; Imperial College London, Centre for Health Policy Greaves, Felix; Imperial College London, Department of Primary Care and Public Health Godwin, Jonathan; DeepMind Technologies Ltd Pearson-Stuttard, Jonathan; Imperial College London, PCPH Harling, Leanne; Imperial College London, Department of Surgery and Cancer McInnes, Matthew; University of Ottawa, Rifai, Nader; Harvard Medical School, Tomasev, Nenad; DeepMind Technologies Ltd Normahani, Pasha; Imperial College London, Department of Surgery and Cancer Whiting, Penny; University of Bristol, School of Social and Community Medicine Aggarwal, Ravi; Imperial College London, Department of Surgery and Cancer Vollmer, Sebastian; The Alan Turing Institute</p>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

	Markar, Sheraz; Imperial College London, Panch, Trishan Liu, Xiaoxuan; University of Birmingham
Primary Subject Heading :	Health informatics
Secondary Subject Heading :	Evidence based practice, Health policy
Keywords :	Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT





I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Developing a Reporting Guideline for Artificial Intelligence Centred Diagnostic Test Accuracy

Studies: The STARD-AI Protocol

Authors:

Viknesh Sounderajah^{1,2}, Hutan Ashrafian^{1,2}, Robert Golub¹¹, Shravya Shetty⁶, Jeffrey De Fauw³, Lotty Hooft¹⁸, Karel Moons¹⁸, Gary Collins¹⁷, David Moher¹², Patrick Bossuyt¹³ and Ara Darzi^{1,2} on behalf of the STARD-AI Steering Committee (Alan Karthikesalingam⁶, Alastair Denniston^{4,14,15,16}, Bilal Mateen¹⁸, Daniel Ting¹⁰, Darren Treanor²⁰, Dominic King²¹, Felix Greaves⁵, Jonathan Godwin³, Jonathan Pearson-Stuttard⁹, Leanne Harling², Matthew McInnes⁷, Nader Rifai²², Nenad Tomasev³, Pasha Normahani², Penny Whiting²³, Ravi Aggarwal¹, Sebastian Vollmer¹⁹, Sheraz Markar², Trishan Panch⁸ and Xiaoxuan Liu^{4,14,15,16})

Author Affiliations

¹ Institute of Global Health Innovation, Imperial College London, United Kingdom

² Department of Surgery and Cancer, Imperial College London, United Kingdom

³ DeepMind, United Kingdom

⁴ Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, United Kingdom

⁵ The National Institute for Health and Care Excellence, United Kingdom

⁶ Google Health

⁷ Department of Radiology, University of Ottawa, Canada

⁸ Division of Health Policy and Management, Harvard T.H. Chan School of Public Health, United States of America

⁹ School of Public Health, Imperial College London, United Kingdom

¹⁰ Singapore Eye Research Institute, Singapore National Eye Center, Singapore

¹¹ JAMA (Journal of the American Medical Association), United States of America

- 1
2
3 27 ¹² Ottawa Hospital Research Institute, Canada
4
5
6 28 ¹³ Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam,
7
8 29 The Netherlands
9
10 30 ¹⁴ University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom
11
12 31 ¹⁵ Health Data Research UK, London, United Kingdom
13
14 32 ¹⁶ Clinical Epidemiology Program, Ottawa Hospital Research Institute, Canada
15
16
17 33 ¹⁷Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and
18
19 34 Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, United Kingdom
20
21 35 ¹⁸Julius Center for Health Sciences and Primary Care, and Cochrane Netherlands, University Medical
22
23 36 Center Utrecht, Utrecht University, Utrecht, The Netherlands
24
25
26 37 ¹⁹ Alan Turing Institute, Kings Cross, United Kingdom
27
28 38 ²⁰ Leeds Teaching Hospitals NHS Trust, University of Leeds, Leeds, United Kingdom
29
30 39 ²¹ Optum, Paddington, London, United Kingdom
31
32 40 ²² Department of Laboratory Medicine, Boston Children's Hospital, Harvard Medical School, Boston,
33
34 41 Massachusetts, United States of America
35
36
37 42 ²³ School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Author disclosures:

The views and opinions expressed herein are those of the authors and do not necessarily reflect the views of their employers or funders.

Corresponding author:

Mr Hutan Ashrafian BSc (Hons) MBBS MRCS PhD MBA

Institute of Global Health Innovation, 10th Floor, Queen Elizabeth Queen Mother building, St Mary's Hospital Campus, Praed Street, London, United Kingdom, W2 1NY

Telephone Number: +447799871597

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

53 **E-mail:** hutan@imperial.ac.uk

54

55

56

57

58

59

60

61

62

63

64

65

66

67 **Data Statement:**

68 There is no data in this work.

69

70 **Word count (main body):**

71 3360

72

73 **Study Status:**

74 Stages 1 and 2 of this study has been completed. Stage 3 is underway (the study is currently between

75 round 1 and round 2 of the modified Delphi consensus process).

76

77

1
2
3 78 **Abstract**
4

5 79 Introduction:
6

7
8 80 STARD was developed to improve the completeness and transparency of reporting in studies
9
10 81 investigating diagnostic test accuracy. However, its current form, STARD 2015 does not address the
11
12 82 issues and challenges raised by artificial intelligence (AI) centred interventions. As such, we propose
13
14 83 an AI-specific version of the STARD checklist (STARD-AI), which focuses upon the reporting of AI
15
16 84 diagnostic test accuracy studies. This paper describes the methods that will be used to develop STARD-
17
18 85 AI.
19

20
21 86
22
23 87 Methods and analysis:
24

25 88 The development of the STARD-AI checklist can be distilled into six stages. (1) A project organisation
26
27 89 phase has been undertaken, during which a Project Team and a Steering Committee were established.
28
29 90 (2) An item generation process has been completed following a literature review, a patient and public
30
31 91 involvement and engagement (PPIE) exercise and an online scoping survey of international experts.
32
33 92 (3) A three-round modified Delphi consensus methodology is underway, which will culminate in a
34
35 93 teleconference consensus meeting of experts. (4) Thereafter, the Project Team will draft the initial
36
37 94 STARD-AI checklist and the accompanying documents. (5) A piloting phase amongst expert users will
38
39 95 be undertaken to identify items which are either unclear or missing. This process, consisting of surveys
40
41 96 and semi-structured interviews, will contribute towards the explanation and elaboration document.
42
43 97 (6) Upon finalisation of the manuscripts, the group's efforts turn towards an organised dissemination
44
45 98 and implementation strategy to maximise end-user adoption.
46
47 99

50
51
52 100 Ethics and dissemination:
53

54 101 Ethical approval has been granted by the Joint Research Compliance Office at Imperial College London
55
56 102 (reference number: 19IC5679). A dissemination strategy will be aimed towards 5 groups of
57
58
59
60

1
2
3 103 stakeholders: (a) academia, (b) policy, (c) guidelines and regulation, (d) industry and (e) public and
4
5 104 non-specific stakeholders. We anticipate that dissemination will take place in Q3 of 2021.
6
7
8 105

9
10 106 Key words:
11

12 107 Diagnostic accuracy, reporting guideline, artificial intelligence, STARD, transparency
13
14 108

15
16 109 Word count: 285/300
17
18
19 110
20
21 111
22
23 112
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

1
2
3 113 **Article Summary**
4

5 114 **Strengths and limitations of this study:**
6

- 7 115
- 8 • Gap: There are no specific reporting standards for artificial intelligence (AI) diagnostic test
9 accuracy studies
10 116
 - 11 117 • Solution: We are developing a specific set of reporting standards for AI diagnostic test
12 accuracy studies; STARD-AI.
13 118
 - 14 119 • Clinical implications: This will help key stakeholders to appraise quality and compare
15 diagnostic test accuracy of AI models that are reported in scientific studies.
16 120
 - 17 121 • Strengths: STARD-AI will be the product of an extensive evidence generation process that is
18 led by multiple stakeholders (clinician scientists, computer scientists, journal editors,
19 EQUATOR Network representatives, reporting guideline developers, epidemiologists,
20 statisticians, industry leaders, funders, health policy makers, patients, legal experts, and
21 medical ethicists).
22 122
 - 23 123 • Limitations: Views of Delphi panellists may differ from those experts who decline
24 participation.
25 124
26 125
27 126
28 127
29 128
- 30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

only

1
2
3 129 **Glossary**
4

5 130 Project Team
6

7 131 This consists of the founder of STARD (PMB), the former United Kingdom Minister for Health and the
8
9 132 current chair for the National Health Service Accelerated Access Collaborative (AD), members of the
10
11 133 TRIPOD-AI group (GSC, KGM), a senior software engineer (SS), directors of the EQUATOR Network
12
13 134 (DM, GSC), the scientific content deputy editor for JAMA (RG) as well as 2 clinician scientists from
14
15 135 Imperial College London (HA, VS).
16
17

18
19 136

20
21 137 Steering Committee
22

23 138 This consists of clinician scientists, computer scientists, journal editors, EQUATOR Network
24
25 139 representatives, epidemiologists, statisticians, industry leaders, funders, health policy makers, legal
26
27 140 experts, and medical ethicists.
28
29

30 141

31
32 142 Consensus Group
33

34 143 This consists of experts who participated in the modified Delphi consensus process (stage 3) of the
35
36 144 study.
37
38

39 145

40
41 146 Pilot Group
42

43 147 This consists of experts who participated in the pilot phase (Stage 5) of the study.
44
45
46 148

47
48 149 Checklist
49

50 150 A document listing the minimally essential items that should be reported in all diagnostic test accuracy
51
52 151 studies centred around artificial intelligence centred index tests. This constitutes the core of the
53
54 152 reporting guideline.
55
56

57 153

58
59 154 Statement
60

1
2
3 155 A document which provides the rationale underpinning the reporting guideline and describes the
4
5 156 process of developing the associated documents.
6
7
8 157

9
10 158 Explanation and Elaboration (E&E)
11

12 159 A document which provides the rationale behind each item in the checklist alongside examples of
13
14 160 good reporting.
15
16 161

17
18 162 Reporting guideline
19

20
21 163 The combination of the checklist, statement and E&E documents.
22
23 164

24
25 165 Artificial Intelligence (AI)
26

27
28 166 The science of developing computer systems which can perform tasks which normally require
29
30 167 human intelligence.
31
32 168

33
34 169 Modified Delphi study
35

36 170 A research method that derives the collective opinions of a group through a staged consultation of
37
38 171 surveys, questionnaires, or interviews, with an aim to reach consensus at the end.
39
40 172

41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

173 **Introduction**

174 Artificial intelligence (AI) is commonly cited as an imminent disruptive innovation[1] within the health
175 sector. If used successfully, AI has the potential to tackle (1) the high rate of avoidable medical errors,
176 (2) workflow inefficiencies and (3) delivery inefficiencies associated with modern healthcare
177 provision[2]. The majority of AI interventions that are close to translation are in the field of medical
178 diagnostics[3]. In the current paradigm, diagnostic investigations require timely interpretation from
179 an expert clinician in order to generate a diagnosis and to subsequently direct episodes of care.
180 However, the recurring issue with the present system is that diagnostic services are inundated with
181 large volumes of work, which often exceeds workforce capacity[4]; COVID-19 being an immediate case
182 in point. In order to address this, diagnostic AI algorithms have positioned themselves as medical
183 devices that may achieve diagnostic accuracy comparable to that of an expert clinician whilst
184 concurrently alleviating health-resource use. Although this paradigm shift may seem imminent, it is
185 crucial to note that much of the evidence supporting diagnostic algorithms has been disseminated in
186 the absence of AI-specific reporting guidelines. Without this guidance, and in a relatively nascent area,
187 key stakeholders are poorly placed to appraise quality and compare diagnostic accuracy between
188 scientific studies.

189 The STARD (Standards for Reporting of Diagnostic Accuracy Studies) 2015 statement remains the most
190 widely accepted set of reporting standards for diagnostic test accuracy studies[5]. STARD was
191 developed to improve the completeness and transparency of studies investigating diagnostic test
192 accuracy. It consists of a checklist of 30 items that authors are strongly encouraged to address when
193 reporting their diagnostic test accuracy studies. It is endorsed by over 200 biomedical journals[6] and
194 studies have shown that adherence to the STARD checklist leads to improved reporting of key study
195 parameters[7,8].

1
2
3 196 However, in its current iteration, STARD 2015 is not designed to address the issues and challenges
4
5 197 raised by AI-driven modalities. Issues include unclear methodological interpretation (e.g., data pre-
6
7 198 processing steps, model development choices and the use of external validation datasets), the lack of
8
9
10 199 standardized nomenclature (e.g., the varying definition of the term 'validation'), as well as the use of
11
12 200 unfamiliar outcome measures (e.g., Jaccard similarity coefficient and F-score). Until these issues are
13
14 201 addressed, achieving comprehensive evaluations of these technologies and their potential
15
16 202 translational benefits will remain limited.

17
18
19
20 203 In order to tackle these problems, we propose an AI-specific STARD guideline (STARD-AI) that aims to
21
22 204 focus upon the reporting of AI diagnostic test accuracy studies[9]. This work is complementary to the
23
24 205 other AI centred checklists listed in the EQUATOR (Enhancing Quality and Transparency of Health
25
26 206 Research) Network program (www.equator-network.org)[10], such as SPIRIT-AI (Standard Protocol
27
28 207 Items: Recommendations for Interventional Trials)[11], CONSORT-AI (Consolidated Standards of
29
30 208 Reporting Trials)[12] and TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for
31
32 209 Individual Prognosis or Diagnosis)[13].

33
34
35
36
37
38 210 STARD-AI is being coordinated by a global Project Team and Steering Committee consisting of clinician
39
40 211 scientists, computer scientists, journal editors, EQUATOR Network representatives, reporting
41
42 212 guideline developers, epidemiologists, statisticians, industry leaders, funders, health policy makers,
43
44 213 legal experts and medical ethicists.

45 46 47 48 214 Aim

49
50
51
52 215 This study aims to produce a specific reporting guideline (STARD-AI) for AI-centred diagnostic test
53
54 216 accuracy studies.

55 56 57 58 217 Focus of STARD-AI

1
2
3 218 The focus of STARD-AI is to aid the comprehensive reporting of research that use AI techniques to
4
5 219 assess diagnostic test accuracy and performance. This can account for either single or combined test
6
7 220 data, which often consists of either (1) imaging data (e.g., CT scans), (2) pathological data (e.g. digitised
8
9
10 221 specimen slide) or (3) reporting data (e.g. electronic health records). STARD-AI may also be used within
11
12 222 studies which report upon image segmentation and other relevant data classification techniques. If
13
14 223 the emphasis of the study is on either developing, validating or updating a multivariable prediction
15
16 224 model which produces an individualised probability of developing a condition (e.g., time-to-event
17
18 225 prediction), the TRIPOD-AI reporting guidelines (Transparent Reporting of a multivariable prediction
19
20 226 model for Individual Prognosis Or Diagnosis) may be more appropriate.
21
22

23 227
24
25 228 Typically, diagnostic test accuracy studies compare test results between participants who are either
26
27 229 with or without a target condition. Data from study participants undergo assessment by an index test,
28
29 230 which is designed to identify a specific target condition. This process occurs alongside a concurrent
30
31 231 reference standard for the target condition within a defined timeframe. Estimates of performance are
32
33 232 typically based on a comparison between index test results and reference standard results from the
34
35 233 same participant cohort. Alternatively, diagnostic performance can compare the performance of an
36
37 234 index test against a reference standard determined through the incidence of an event within a defined
38
39 235 timeframe.
40
41
42

43 236
44
45 237 A significant number of contemporary AI diagnostic studies include information related to both the
46
47 238 development and testing (validation) of AI centred index tests. In order to accommodate and improve
48
49 239 upon this practice, STARD-AI will propose items related to AI index test development and validation
50
51 240 as part of the consensus process. Other key topics for consideration within this study include, but are
52
53 241 not limited to, the following: (1) data pre-processing methods, (2) AI index test development methods
54
55 242 (e.g., dataset partition, model calibration, stopping criteria when training, use of external validation
56
57 243 sets), (3) fairness metrics, (5) non-standard performance metrics, (5) explainability and (6) human-AI
58
59
60

1
2
3 244 index test interaction. As noted in the methods section, the inclusion of specific items related to these
4
5 245 issues is reliant upon consensus that is achieved through a transparent and fair evidence generation
6
7
8 246 process.
9

10 247
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

248 **Methods**

249 This protocol has been constructed in accordance with the EQUATOR Network (Enhancing the Quality
250 and Transparency of Health Research) toolkit for developing reporting guidelines[14]. It has also
251 greatly benefitted from the experience and expertise from Project Team and Steering Committee
252 members who had previously led the STARD 2003[15], STARD 2015, STARD for Abstracts[16], SPIRIT-
253 AI and CONSORT-AI initiatives respectively.

254 We can distil the development of the STARD-AI checklist into six stages. The overall goal of the STARD-
255 AI initiative is to generate a list of minimally essential items, based upon the established STARD 2015
256 framework, that should be reported in all AI diagnostic test accuracy studies. The items must assist
257 the reader to appraise the completeness, applicability, and potential for bias of the study findings.

258 Stage 1: Project organisation

259 A nine member STARD-AI Project Team was established to coordinate the reporting guideline
260 development process. The Project Team consists of the founder of STARD (PMB), the former United
261 Kingdom Minister for Health and the current chair for the National Health Service Accelerated Access
262 Collaborative (AD), members of the TRIPOD-AI core committee (GSC, KGM), a senior software
263 engineer (SS), directors of the EQUATOR Network (DM, GSC), the scientific content deputy editor for
264 JAMA (RG) as well as 2 clinician scientists from Imperial College London (HA, VS). The Project Team
265 are responsible for identifying suitable members of the Steering Committee, candidate item
266 generation, undertaking the online surveys for the modified Delphi consensus process, organising the
267 consensus meeting, drafting the STARD-AI checklist and accompanying documents, piloting the draft
268 STARD-AI checklist as well as leading upon the dissemination process.

1
2
3 269 Further to the Project Team, a multidisciplinary STARD-AI Steering Committee was established to
4
5 270 provide specialist guidance throughout. This committee consists of clinician scientists, computer
6
7 271 scientists, journal editors, EQUATOR network directors, epidemiologists, statisticians, industry
8
9 272 leaders, funders, health policy leaders, regulatory leaders, legal experts, patient representation
10
11 273 experts and medical ethicists. These individuals were identified through their notable work with
12
13 274 respect to (1) diagnostic accuracy research and its clinical translation, (2) applied artificial intelligence
14
15 275 in healthcare as well as (3) notable contribution to other AI-centred EQUATOR Network registered
16
17 276 initiatives, such as TRIPOD-AI, CONSORT-AI and SPIRIT-AI.
18
19
20
21
22

23 277 Prior to Stage 2, the STARD-AI project was registered with the EQUATOR Network.
24
25
26

27 278 Stage 2: Item generation
28
29

30
31 279 In order to generate a candidate list of items to enter the modified Delphi consensus process, the
32
33 280 Project Team undertook a literature review, an online scoping survey with an international panel of
34
35 281 experts and a patient public involvement and engagement (PPIE) exercise.
36
37
38

39 282 a) Literature review:
40
41

42 283
43
44 284 In January 2020, a literature review of both academic and non-academic literature was undertaken.
45
46 285 An electronic database search of Medical Literature Analysis and Retrieval System Online (MEDLINE)
47
48 286 and Excerpta Medica database (EMBASE) was conducted through Ovid. Both Medical Subject Headings
49
50 287 (MeSH) or EMBASE Subject Headings (Emtree) were used. Search results were imported into
51
52 288 Covidence (Covidence.org, Melbourne, Australia) for duplicate removal and study selection. Two
53
54 289 individuals (VS,HA) individually screened study titles and abstracts for inclusion. Disagreements were
55
56 290 resolved through discussion.
57
58
59
60

1
2
3 292 This process was augmented by non-systematic searches using grey literature, social networking
4
5 293 platforms as well as personal article collections highlighted by members of the Project Team. Titles
6
7 294 and abstracts of shortlisted publications were screened by one of two reviewers (VS, HA) and
8
9
10 295 potentially eligible publications were retrieved for full-text assessment. Extracted material were
11
12 296 broadly classified into four categories: (1) general considerations regarding diagnostic accuracy
13
14 297 studies and artificial intelligence, (2) evidence and statements suggesting modification to existing
15
16 298 STARD 2015 items, (3) evidence and statements suggesting additions to the STARD 2015 checklist and
17
18 299 (4) evidence and statements suggesting the removal of specific items from the STARD 2015 checklist.
19
20

21 300

22
23 301 b) Online scoping survey:

24
25
26
27 302 In addition to this, in February 2020, the Project Team undertook an online survey with an
28
29 303 international panel of 80 experts in order to identify potential further items or modifications that
30
31 304 warrant consideration. Written participant consent was attained as part of this process. This process
32
33 305 generated over 2500 responses, which were analysed and classed into the aforementioned 4 broad
34
35 306 categories.
36
37
38

39
40 307 c) Patient public involvement and engagement (PPIE) exercise:

41 308

42
43
44
45 309 Lastly, a focus group was conducted with patients and members of the public who had expressed an
46
47 310 interest in participating in forums related to digital health and AI. Written participant consent was
48
49 311 attained as part of this process. The objective of these discussions was two-fold; (1) to further identify
50
51 312 issues not uncovered during the literature review and expert survey and (2) to gain further
52
53 313 understanding of the perceived importance of specific items raised thus far. These discussions were
54
55 314 conducted remotely using Zoom (Zoom Video Communications, Inc., USA).
56
57

58 315

1
2
3 316 An expert facilitator led a discussion on the current use of AI in healthcare, on what the aims of STARD-
4
5 317 AI were and what participants considered to be important items to capture during the study process.
6
7 318 As stakeholder discussions were conducted virtually on Zoom, anonymised post-hoc discussion
8
9 319 transcripts were maintained. Two investigators (VS, HA) independently identified common themes
10
11 320 and sub-themes from the discussion, which were classed into the aforementioned 4 broad categories.
12
13
14 321
15
16 322 Having synthesised the findings of the literature review, the survey and the patient public involvement
17
18 323 and engagement exercise, the Project Team, in collaboration with the Steering Committee, decided
19
20 324 upon which items warranted consideration in the formal modified Delphi consensus process.
21
22
23
24

25 325 Stage 3: Modified Delphi consensus process (ongoing)

26
27
28
29 326 a) Study design and participants:
30
31 327

32
33 328 This study has adopted a pragmatic modified Delphi consensus methodology. The Delphi consensus
34
35 329 methodology is a well-established method[17] of obtaining a collective opinion from a group of
36
37 330 experts through a series of questionnaires; each one refined based upon feedback from respondents.
38
39
40 331
41
42 332 Participants were invited to join the STARD-AI Consensus Group on account of their expertise as
43
44 333 clinician scientists, computer scientists, journal editors, EQUATOR Network representatives, reporting
45
46 334 guideline developers, epidemiologists, statisticians, industry leaders (e.g., clinician scientists,
47
48 335 computer scientists and product managers from health technological companies), funders, health
49
50 336 policy makers, legal experts and medical ethicists. These experts were shortlisted through two
51
52 337 principle means; either through the professional networks of members of the STARD-AI Project Team
53
54 338 and Steering Committee or through recognition, critical involvement and achievements in a field that
55
56 339 is related to diagnostic AI systems in the health sector (e.g., authorship of seminal academic
57
58
59
60

1
2
3 340 publications, key thought leaders, clinicians involved in prominent AI translational work and health
4
5 341 policy directors, amongst others). Moreover, ensuring fair representation across geographies and
6
7 342 demographics was a pertinent consideration during recruitment. Shortlisted participants were
8
9
10 343 mutually agreed upon by the Project Team members.
11

12 344

14 345 Following this, invited experts were provided with three weeks to respond to the initial invitation to
15
16 346 participate. Written participant consent was attained as part of this process. Those who accepted the
17
18 347 invitation were invited to complete each round of the modified Delphi consensus process. Those who
19
20 348 contribute to both online rounds will be acknowledged by name as an author, within a group
21
22 349 authorship model, in the publication that arises from this study.
23

24 350

27 351 In each round of the modified Delphi consensus process, participants are asked to grade each
28
29 352 candidate item using a 5-point Likert-like scale (1 – very important, 2 – important, 3 – moderately
30
31 353 important, 4 – slightly important, 5 – not at all important). The threshold for consensus is predefined
32
33 354 at $\geq 75\%$. Items which achieve $\geq 75\%$ ratings of 1 or 2 are deemed to be essential for inclusion and are
34
35 355 put forward for discussion in the final round (round 3, which will occur in the form of a virtual
36
37 356 teleconference meeting). Items which achieve $\geq 75\%$ ratings of 4 or 5 are deemed unimportant for
38
39 357 inclusion and are excluded. Items which do not reach this threshold of consensus are put forward to
40
41 358 the next round of the modified Delphi consensus process. In addition to rating items, participants are
42
43 359 asked in a free-text format to suggest any other items that they consider to be important to discuss in
44
45 360 subsequent rounds.
46
47

48 361

50
51
52 362 In round 2, the survey will compose of (1) items for which consensus was not achieved in round 1 and
53
54 363 (2) any new items suggested as part of round 1 feedback. Next to each item, participants will be
55
56 364 reminded of what rating they gave in the previous round. Additionally, the mean score given by the
57
58 365 overall group in the previous round will be displayed for each item. Thus, participants will be able to
59
60

1
2
3 366 revise their initial score with the additional knowledge of peer responses. Following the collection of
4
5 367 round 2 responses, additional items which achieve consensus as 'important' will be put forward for
6
7 368 discussion during round 3. Those items that achieve consensus as 'unimportant' are excluded. Lastly,
8
9 369 any non-consensus items from round 2 will be resolved through discussion amongst those in virtual
10
11 370 attendance at the consensus meeting (round 3).
12
13

14 371

15
16 372 b) Round 3; the consensus meeting:
17
18

19 373

20
21 374 The consensus meeting (round 3) will consist of the STARD-AI Project Team and the STARD-AI Steering
22
23 375 Committee. Given COVID-19 constraints, the meeting will be conducted virtually using Zoom. The
24
25 376 primary objective is to develop a draft version of the STARD-AI checklist. As recommended in the
26
27 377 COMET handbook, the nominal group technique, a highly-structured group interaction framework,
28
29 378 will be utilised to aid this process[18,19]. Following a brief introduction and explanation of the purpose
30
31 379 of the meeting by the facilitators (VS, HA), participants will discuss the inclusion and exclusion of
32
33 380 candidate items. Participants will be asked to share any comments they have generated in a 'round
34
35 381 robin' format until all contributions are exhausted. Participants will then be invited to discuss or seek
36
37 382 further clarification about any of the ideas or comments produced. This discussion phase will be led
38
39 383 by facilitators (VS, HA) to ensure that the discussion will not be dominated by any one individual and
40
41 384 will be as neutral as possible[20].
42
43
44
45

46 385

47
48 386 c) Study conduct:
49

50 387

51
52 388 VS and HA are the Delphi facilitators for the online survey rounds as well as the teleconference
53
54 389 consensus meeting. They are responsible for the creation of the questionnaires, the invitations, the
55
56 390 responses, the reminders, the analysis as well as the feedback for subsequent rounds.
57
58
59

60 391

1
2
3 392 The first two rounds of the modified Delphi consensus process are conducted as online surveys using
4
5 393 the DelphiManager software (version 4.0), which is developed and maintained by the COMET (Core
6
7 394 Outcome Measures in Effectiveness Trials) initiative. Round 3 (the consensus meeting) will be carried
8
9
10 395 using Zoom.

11
12 396

13
14 397 Stage 4: Development of the (1) checklist, (2) statement and (3) explanation and elaboration (E&E)
15
16 398 document

17
18
19
20 399 Upon completion of the modified Delphi consensus process, the Project Team will draft the initial
21
22 400 STARD-AI checklist and statement. The draft checklist and statement will be shared amongst the wider
23
24 401 Steering Committee in order to discuss its content and therefore allow the Steering Committee to
25
26 402 suggest additions, subtractions or modifications as they see fit. This stage will also allow for
27
28 403 harmonisation of key terms with the imminent TRIPOD-AI, in addition to the existing CONSORT-AI and
29
30 404 SPIRIT-AI checklists.

31
32
33
34
35 405 Stage 5: Piloting phase

36
37
38
39 406 Upon completion of the first draft of the STARD-AI checklist, we intend to organise a piloting phase
40
41 407 amongst expert users (Pilot Group). The main aim of these piloting sessions is to identify items which
42
43 408 are considered to be vague, unnecessary or missing. We intend to undertake this process amongst
44
45 409 radiology experts, pathology experts, computer scientists, expert statisticians, journal editorial
46
47 410 boards, members of the global EQUATOR Network, key industry stakeholders as well as policy experts.
48
49 411 Much like stage 3, these experts are shortlisted through two principle means; either through the
50
51 412 professional networks of members of the STARD-AI Project Team and Steering Committee or through
52
53 413 either (1) involvement in teams that have led diagnostic AI studies or (2) work as peer reviewers or
54
55 414 editorial board members for journals that publish diagnostic AI studies. Experts are mutually agreed
56
57 415 upon by the Project Team members and Steering Committee. Feedback will be captured through

1
2
3 416 surveys and a series of semi-structured interviews. This approach allows for the capture of broad
4
5 417 issues through surveys, which form themes that can be further explored in detail during semi-
6
7 418 structured interviews. Anonymised feedback from the interviews will be transcribed to allow for
8
9 419 thematic analysis so that recurring trends are appropriately identified and presented back to the
10
11 420 Project Team and Steering Committee for discussion. Experts within the Pilot Group will be
12
13 421 acknowledged by name as an author, within a group authorship model, in the publications that arise
14
15 422 from this study.
16
17
18

19 423
20
21 424 In conjunction to this piloting process, the Project Team will also prepare the explanation and
22
23 425 elaboration (E&E) document to provide rationale for the included items alongside examples of good
24
25 426 reporting.
26
27
28

29 427 Stage 6: Finalisation, publication, and post-publication activities

30
31
32
33 428 Following the piloting phase, the final proposed amendments to STARD-AI will be discussed amongst
34
35 429 the Project Team and the Steering Committee. Once consensus has been reached through e-mail
36
37 430 correspondence, the checklist and accompanying documents will be disseminated.
38
39
40

41
42 431 The dissemination strategy will be principally tailored towards 5 groups of stakeholders; (a) academia,
43
44 432 (b) policy, (c) guidelines and regulation, (d) industry and (e) patient representing bodies. Although a
45
46 433 significant amount of material will cross over between stakeholders, creating specific material is
47
48 434 considered to be the most meaningful way of achieving impact.
49
50

51
52 435 We aim to publish the STARD-AI checklist, the accompanying statement and the E&E document in an
53
54 436 open access format (through a CC-BY license). In order to further complement this, we aim to create
55
56 437 specialty-specific discourse regarding STARD-AI through focussed editorials in pertinent journals.
57
58 438 These journal editors will also be actively encouraged to endorse STARD-AI as part of their broader
59
60

1
2
3 439 editorial policy. Moreover, we will present STARD-AI at national and international scientific meetings.
4
5 440 Translations of the guideline in various languages are actively encouraged (available on the EQUATOR
6
7 441 network) in order to further broaden the scope of its impact. We encourage interested parties to
8
9 442 contact the corresponding author for further information about the translation policies.
10
11
12

13
14 443 In addition to this, we aim to persuade governmental bodies to adopt the checklist as part of their
15
16 444 policy assessments. This will involve presentations at national and international health policy summits
17
18 445 (e.g., World Innovation Summit for Health and NHS Accelerated Access Collaborative meetings).
19
20 446 Furthermore, we will aim to integrate teaching about STARD-AI into national health policy educational
21
22 447 programmes through pre-existing collaborations with academic institutions, NHS Digital Academy and
23
24 448 NHSX.
25
26
27

28
29 449 Concurrent to this workstream will be our work with guidelines and regulatory bodies so that they
30
31 450 may account for STARD-AI as part of their national health technology assessments. This will involve
32
33 451 the United States Food and Drug Administration (FDA), the Medicines and Healthcare products
34
35 452 Regulatory Agency (MHRA) and The National Institute for Health and Care Excellence (NICE) amongst
36
37 453 others.
38
39
40

41
42 454 Lastly, we will present STARD-AI to a broad range of health technology companies so that their product
43
44 455 pipelines may accommodate for this downstream mode of assessment.
45
46
47

48 456 Conclusion:

49
50
51
52 457 STARD-AI will serve as the first global-consensus achieved guidance for the reporting of AI centred
53
54 458 diagnostic accuracy studies. Through a clear multi-stakeholder dissemination policy, we hope that
55
56 459 STARD-AI can significantly contribute towards minimising research waste as well as serving as an
57
58
59
60

1
2
3 460 instrument that assists the streamlined translation of these nascent technologies. We anticipate that
4
5 461 STARD-AI will be published in Q3 2021.
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

1
2
3 463 **Ethics**
4
5
6

7 464 Ethical approval has been granted by the Joint Research Compliance Office at Imperial College
8
9 465 London (SETREC reference number: 19IC5679).
10

11 466
12

13
14 467 **Author Statement**
15

16 468 Viknesh Sounderajah, Hutan Ashrafian, Robert Golub, Shravya Shetty, Jeffrey De Fauw, Lotty Hooft,
17
18 469 Karel Moons, Gary Collins, David Moher, Patrick Bossuyt and Ara Darzi were involved in the planning
19
20 470 and design of the study. Viknesh drafted the manuscript with all authors contributing to the writing.
21
22

23 471
24

25 472 Alan Karthikesalingam, Alastair Denniston, Bilal Mateen, Daniel Ting, Darren Treanor, Dominic King,
26
27 473 Felix Greaves, Jonathan Godwin, Jonathan Pearson-Stuttard, Leanne Harling, Matthew McInnes,
28
29 474 Nader Rifai, Nenad Tomasev, Pasha Normahani, Penny Whiting, Ravi Aggarwal, Sebastian Vollmer,
30
31 475 Sheraz Markar, Trishan Panch and Xiaoxuan Liu are members of the STARD-AI Steering Committee.
32
33 476 They are equally involved in the wider conduct and direction of the overall study. All of the authors
34
35 477 edited the manuscript and provided critical appraisal.
36
37

38 478
39

40 479 All named authors approved the final draft of the manuscript.
41
42

43 480
44

45 481 **Competing Interests**
46

47 482 There are no competing interests for any author.
48
49

50 483
51

52 484 **Funding**
53

54 485 Infrastructure support for this research was provided by the NIHR Imperial Biomedical Research
55
56 486 Centre (BRC).
57
58
59
60

1
2
3 487 GSC is supported by the NIHR Oxford Biomedical Research Centre and Cancer Research UK
4
5 488 (programme grant: C49297/A27294).
6

7 489 DT is funded by National Pathology Imaging Co-operative, NPIC (Project no. 104687) is supported by
8
9 490 a £50m investment from the Data to Early Diagnosis and Precision Medicine strand of the
10
11 491 government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and
12
13 492 Innovation (UKRI).
14
15

16 493 FG is supported by the National Institute for Health Research Applied Research Collaboration
17
18 494 Northwest London
19
20

21 495
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

496 **References**

- 497 1 Sounderajah V, Patel V, Varatharajan L, *et al.* Are disruptive innovations recognised in the
498 healthcare literature? A systematic review. *BMJ Innov* 2020;:bmjinnov-2020-000424.
499 doi:10.1136/bmjinnov-2020-000424
- 500 2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence.
501 *Nat. Med.* 2019;**25**:44–56. doi:10.1038/s41591-018-0300-7
- 502 3 Benjamens S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved
503 medical devices and algorithms: an online database. *npj Digit Med* 2020;**3**:118.
504 doi:10.1038/s41746-020-00324-0
- 505 4 Williams BJ, Bottoms D, Treanor D. Future-proofing pathology: The case for clinical adoption
506 of digital pathology. *J Clin Pathol* 2017;**70**:1010–8. doi:10.1136/jclinpath-2017-204644
- 507 5 Bossuyt PM, Reitsma JB, Bruns DE, *et al.* STARD 2015: An updated list of essential items for
508 reporting diagnostic accuracy studies. *BMJ* 2015;**351**. doi:10.1136/bmj.h5527
- 509 6 Ochodo EA, Bossuyt PM. Reporting the Accuracy of Diagnostic Tests: The STARD Initiative 10
510 Years On. *Clin Chem* 2013;**59**:917–9. doi:10.1373/clinchem.2013.206516
- 511 7 Korevaar DA, Van Enst WA, Spijker R, *et al.* Reporting quality of diagnostic accuracy studies: A
512 systematic review and meta-analysis of investigations on adherence to STARD. *Evid. Based.*
513 *Med.* 2014;**19**:47–54. doi:10.1136/eb-2013-101637
- 514 8 Korevaar DA, Wang J, Van Enst WA, *et al.* Reporting diagnostic accuracy studies: Some
515 improvements after 10 years of STARD. *Radiology* 2015;**274**:781–9.
516 doi:10.1148/radiol.14141160

- 1
2
3 517 9 Sounderajah V, Ashrafian H, Aggarwal R, *et al.* Developing specific reporting guidelines for
4
5 518 diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat.*
6
7 519 *Med.* 2020;**26**:807–8. doi:10.1038/s41591-020-0941-1
8
9
10
11 520 10 The EQUATOR Network | Enhancing the QUALity and Transparency Of Health Research.
12
13 521 <https://www.equator-network.org/> (accessed 26 Sep 2020).
14
15
16
17 522 11 Rivera SC, Liu X, Chan A-W, *et al.* Consensus statement Guidelines for clinical trial protocols
18
19 523 for interventions involving artificial intelligence: the SPIRIT-AI extension The SPIRIT-AI and
20
21 524 CONSORT-AI Working Group*, SPIRIT-AI and CONSORT-AI Steering Group and SPIRIT-AI and
22
23 525 CONSORT-AI Consensus Group. *Nat Med* 2020;**26**:1351–63. doi:10.1038/s41591-020-1037-7
24
25
26
27
28 526 12 Liu X, Rivera SC. Consensus statement Reporting guidelines for clinical trial reports for
29
30 527 interventions involving artificial intelligence: the CONSORT-AI extension^{6,13} and The
31
32 528 SPIRIT-AI and CONSORT-AI Working Group*. *Nat Med* 2020 269 2020;**26**:1364–74.
33
34 529 doi:10.1038/s41591-020-1034-x
35
36
37
38
39 530 13 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet.*
40
41 531 2019;**393**:1577–9. doi:10.1016/S0140-6736(19)30037-6
42
43
44
45 532 14 Toolkits | The EQUATOR Network. <https://www.equator-network.org/toolkits/> (accessed 26
46
47 533 Sep 2020).
48
49
50
51 534 15 Bossuyt PM, Reitsma JB, Bruns DE, *et al.* The STARD statement for reporting studies of
52
53 535 diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;**138**.
54
55 536 doi:10.7326/0003-4819-138-1-200301070-00012-w1
56
57
58
59
60

- 1
2
3 537 16 Cohen JF, Korevaar DA, Gatsonis CA, *et al*. STARD for Abstracts: Essential items for reporting
4
5 538 diagnostic accuracy studies in journal or conference abstracts. *BMJ* 2017;**358**.
6
7 539 doi:10.1136/bmj.j3751
8
9
10
11 540 17 Brown BB. Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts.
12
13 541 Published Online First: 1968.<https://www.rand.org/pubs/papers/P3925.html> (accessed 26
14
15 542 Sep 2020).
16
17
18
19
20 543 18 McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. *Int J*
21
22 544 *Clin Pharm* 2016;**38**:655–62. doi:10.1007/s11096-016-0257-x
23
24
25
26 545 19 Williamson PR, Altman DG, Bagley H, *et al*. The COMET Handbook: version 1.0. *Trials*
27
28 546 2017;**18**:280. doi:10.1186/s13063-017-1978-4
29
30
31
32 547 20 Harvey N, Holmes CA. Nominal group technique: An effective method for obtaining group
33
34 548 consensus. *Int J Nurs Pract* 2012;**18**:188–94. doi:10.1111/j.1440-172X.2012.02017.x
35
36
37
38
39 549
40
41
42
43 550
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60