COMMENTARY

Ann Acad Med Singap 2023;52:199-212
https://doi.org/10.47102/annals-acadmedsg.2022452

Artificial intelligence innovation in healthcare: Relevance of reporting guidelines for clinical translation from bench to bedside

Zhen Ling Teo *[1] *MBBS*, Ann Kwee *[2] *MRCP*, John CW Lim [3] *SM*, Carolyn SP Lam [4,5] *PhD*, Dean Ho [6] *PhD*, Sebastian Maurer-Stroh [7,8] *PhD*, Yi Su [9] *PhD*, Simon Chesterman [10,11] *DPhil*, Tsuhan Chen [11,12] *PhD*, Chorh Chuan Tan [13] *FRCP*, Tien Yin Wong [1,14] *PhD*, Kee Yuan Ngiam [15] *FRCS*, Cher Heng Tan [16] *FRCR*, Danny Soon [17] *MD*, May Ling Choong [18] *MBBS*, Raymond Chua [19] *FAMS*, Sutowo Wong [20] *B.Eng*, Colin Lim [21] *MBA*, Wei Yang Cheong [21] *PhD*, Daniel SW Ting [1,5,22] *PhD*

ABSTRACT

Artificial intelligence (AI) and digital innovation are transforming healthcare. Technologies such as machine learning in image analysis, natural language processing in medical chatbots and electronic medical record extraction have the potential to improve screening, diagnostics and prognostication, leading to precision medicine and preventive health. However, it is crucial to ensure that AI research is conducted with scientific rigour to facilitate clinical implementation. Therefore, reporting guidelines have been developed to standardise and streamline the development and validation of AI technologies in health. This commentary proposes a structured approach to utilise these reporting guidelines for the translation of promising AI techniques from research and development into clinical translation, and eventual widespread implementation from bench to bedside.

Ann Acad Med Singap 2023;52:199-212

Keywords: Artificial intelligence, clinical translation, digital innovation, guidelines

[1] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[2] Department of Endocrinology, Singapore General Hospital, Singapore
[3] Centre of Regulatory Excellence, Duke-NUS Medical School, National University of Singapore, Singapore
[4] Department of Cardiology, National Heart Centre Singapore, Singapore
[5] Duke-NUS Medical School, National University of Singapore, Singapore
[6] Department of Biomedical Engineering, Institute of Digital Medicine, N.1 Institute of Health and Department of Pharmacology, National University of Singapore, Singapore
[7] Bioinformatics Institute and Infectious Diseases Labs, Agency for Science, Technology and Research, Singapore
[8] Yong Loo Lin School of Medicine and Department of Biological Sciences, National University of Singapore, Singapore
[9] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[10] Faculty of Law, National University of Singapore, Singapore
[11] AI Singapore, Singapore
[12] School of Computing, National University of Singapore, Singapore
[13] Chief Health Scientist Office, Ministry of Health, Singapore
[14] Tsinghua Medicine, Tsinghua University, Beijing, China
[15] Group Technology Office, National University Health System, Singapore
[16] Centre for Health Innovation, National Healthcare Group, Singapore
[17] Consortium for Clinical Research and Innovation, Singapore, Singapore
[18] Health Sciences Authority, Singapore
[19] Director of Medical Services Office (Health Regulation Group), Ministry of Health, Singapore
[20] Data Analytics, Ministry of Health, Singapore
[21] Technology, Ministry of Health, Singapore
[22] Artificial Intelligence Office, Singapore Health Services, Singapore
* Joint first authors
Correspondence: Dr Daniel SW Ting, Singapore Eye Research Institute, Singapore National Eye Centre, 11 Third Hospital Avenue, Singapore 168751.
Email: daniel.ting@duke-nus.edu.sg
Dr Wei Yang Cheong, Ministry of Health, 16 College Road, College of Medicine Building, Singapore 169854.
Email: cheong_wei_yang@moh.gov.sg

Artificial intelligence (AI) and digital innovation have revolutionised many sectors and industries, prominently including healthcare during the coronavirus disease 2019 (COVID-19) pandemic.[1] For example, deep learning, which is a subset of the state-of-the-art machine learning techniques, has shown robust performance in image recognition, speech recognition and natural language processing.[2] In healthcare, machine learning and deep learning are now becoming increasingly adopted as part of segmentation, classification and prediction tasks in image analysis,[3,4] including differentiation between benign and malignant lesions in skin photographs,[5] diabetic retinopathy detection on colour fundus photographs,[6] and detection of COVID-19 or tuberculosis from chest imaging.[7,8] In addition, natural language processing has been heavily adopted in medical chatbots using speech or text,[3] and in the extraction of useful information from electronic medical records.[9] Such AI technologies could be applied in diverse clinical settings, ranging from screening, triaging or remote monitoring at community-based or population-based settings, to performing diagnoses or prognostication in tertiary or quaternary personalised medicine workflows. More recently, the use of deep reinforcement learning techniques has proven to be robust in the prediction of protein folding, potentially unravelling an exciting, untapped avenue for drug discovery research in the proteomics space, albeit in its infancy.[10-12] Prospective validation of AI-based drug intervention that is dynamically tailored to each patient is also being observed.[13]

Other global AI trends and advances, in addition to the above examples, include privacy preserving technologies like federated machine learning, blockchain[14-16], synthetic AI[17] and explainable AI.[18] If used appropriately, AI that is applied in health can bring clinicians a step closer to precision medicine and predictive and preventative health, with the promise of population-wide interventions that could potentially increase early accessibility to appropriate care and greater cost-effectiveness.[19]

To realise the promises of AI in healthcare, it is critical to ensure that AI research is performed and reported with the scientific rigor required for clinical implementation. In this regard, clear reporting guidelines are needed. In parallel with these technological innovations, various AI reporting consensus guidelines have been developed over the past few years, including the Consolidated Standards of Reporting Trials (CONSORT)-AI,[20] Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT)-AI,[21] Developmental and Exploratory Clinical Investigations of Decision Support Systems Drive by Artificial Intelligence (DECIDE-AI),[22] Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-AI[23] and Standards for Reporting of Diagnostic Accuracy Study (STARD)-AI.[24] These are helpful in standardizing and streamlining the development and validation of AI technologies in health, but their relevance at different phases of the AI innovation journey is not always clear. This commentary makes the case for the relevance of these reporting guidelines and proposes a structured approach to utilise them for facilitating the translation of promising techniques from research and development into clinical translation and eventual widespread implementation from bench to bedside.

Why the need for reporting guidelines in AI research?

First, a major barrier to widespread clinical implementation is heterogeneity in study methodology and reporting, which impedes the ability of readers (including clinicians) to critically appraise these new AI technologies. Previous systematic reviews and meta-analyses in AI medical imaging suggest that many studies are suboptimally designed and delivered.[25,26] Differences in terminology used as well as differences in AI model statistics further impede the clinicians' ability to reproduce the results, compare different studies and weigh the evidence for clinical use. Second, AI algorithms perform best in the artificial environment in which they are trained and validated, and generalisability of the algorithms may be limited, particularly where there are significant differences in actual clinical workflows across unique settings in which the software platforms or data exchange protocols are different.[26] Efforts to include heterogenous datasets and demonstrate generalisability via external validation sets are making headway. Third, despite promising explainability research looking into the "black box" of certain AI decision-making processes, the lack of clearly defined intended use and users during the creation of AI models can result in excellent AI models failing to find a role in actual clinical practice.[27] Importantly, the mode in which data is acquired versus the quantity of data acquired is a critical consideration that will also impact the actionability and efficacy of the downstream model. For example, in AI-guided clinical decision support, the longitudinal response of a patient exposed to variable dosing (which can provide a comprehensive picture of the patient's potential for favourable treatment response) is assessed, as opposed to assessing a fixed dose and fixed timepoint, which represents only a snapshot of their response. Thus, AI-guided clinical decision support can lead to substantially different treatment options and outcomes.[28,29] In such an instance, the clinicians themselves play a role in dataset building that can, in turn, align the decision-making processes of the model at the point of care.

To improve clinical translation and acceptance of AI technologies, researchers need to clearly define the intended use, show transparency in their methodology to allow for reproducibility, demonstrate clinical safety, and clearly state the role of the AI technology in current clinical practice. In view of these concerns, several international parties have banded together to develop guidelines to promote transparency and completeness in conducting and in the reporting of AI-related research. The Enhancing the Quality and Transparency of Health Research (EQUATOR) Network is an international initiative that aims to improve the quality of healthcare research by promoting the development and use of robust reporting guidelines. It aims to achieve accurate, complete and transparent reporting of all health research studies to support reproducibility, validity and usefulness.[30]

Overview of existing AI reporting guidelines

Reporting guidelines were developed to specify the minimum information required in published scientific papers and to aid editors, peer reviewers and, most importantly, general readers including clinicians in appraising the quality, value and clinical relevance of these studies. With the recent rise in AI health research, several AI reporting guidelines, some of which serve as an extension of existing medical reporting guidelines, have been proposed to better suit AI studies, which often encompass specific technical details, AI terminology and statistical evaluation that differ from other scientific research. We provide an overview of these guidelines to aid clinicians in appraising AI research and development outcomes for consideration of clinical real-life adoption, aside from their use in the research setting. For the initial research and development phase, STARD-AI[24] and the Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis or Diagnosis (TRIPOD)-AI[31] are useful reporting guidelines to provide a comprehensive framework to develop and test the diagnostic performance of AI algorithms using retrospective data; whereas CONSORT-AI[20] and SPIRIT-AI[21] are mainly for those AI algorithms that have been developed with an established operating threshold and are ready to be tested on a prospective clinical trial dataset. DECIDE-AI was designed to guide early stages of clinical evaluation and analyse the human factors to improve clinical translation of AI studies.[22]

AI diagnostic accuracy testing studies: STARD-AI and QUADAS-AI

STARD-AI is an extension of the STARD guidelines specifically for AI diagnostic test accuracy studies.[24] STARD-AI can be used on a variety of data types, including imaging data, and would be more suitable in the initial research and development stage when retrospective data are used. For diagnostic studies, it is important to specify the intended use environment, set the operating thresholds using preset sensitivity and specificity, and test AI algorithms on independent local and international datasets to illustrate generalisability. Complementary to STARD-AI is QUADAS-AI, which aims to serve as a tool to assess bias and applicability of diagnostic AI systems.[23] STARD-AI and QUADAS-AI are currently under development and, once available, will allow end users to appraise the quality of the AI diagnostic test and allow for comparisons of diagnostic accuracy between studies. A case example will be the use of STARD-AI and QUADAS-AI in a study to evaluate the diagnostic accuracy of a deep learning model for diabetic retinopathy detection on fundus photography.

AI clinical prediction models: TRIPOD-AI and PROBAST-AI

With guidance from the EQUATOR Network, TRIPOD-AI and the Prediction Model Risk of Bias Assessment Tool (PROBAST)-AI are intended as reporting guidelines for studies on AI machine learning-based prediction models and their risk assessment.[31] TRIPOD focuses on improving the transparency of a prediction model irrespective of the methodology,[32] whereas PROBAST puts heavy emphasis on evaluating the risk of bias and applicability of the primary prediction model, especially for a systematic review.[33] Both AI extensions are currently under development and will eventually aid researchers and clinicians to critically appraise machine learning prediction models and provide a standardised tool for the evaluation of study bias among AI studies. For example, TRIPOD-AI or PROBAST-AI will be suitable for a study on a machine learning model that can predict the probability of future cardiac arrest based on clinical and imaging data.

AI clinical trials: SPIRIT-AI and CONSORT-AI

With the support of the EQUATOR Network, SPIRIT and CONSORT are guidelines that are now widely accepted as international standards.[34,35] Further extensions for clinical trials specifically involving AI were published in 2020 and are known as SPIRIT-AI[21] and CONSORT-AI.[20]

SPIRIT-AI and CONSORT-AI were designed by an international multistakeholder group amid mounting recognition that interventions involving AI require rigorous evaluation to prove their impact on health outcomes. Extensions of existing guidelines are required to assist editors, peer reviewers and general readership to understand and critically appraise these AI-related interventions. SPIRIT-AI is a guideline for AI clinical trial protocols, while CONSORT-AI is a guideline for reporting randomised trials using AI (Table 1).

Briefly, SPIRIT-AI included 15 additional items (3 elaborations of existing items in SPIRIT and 12 new AI extensions) and CONSORT-AI included 14 additional items (3 elaborations of existing items in CONSORT and 11 new AI extensions). Proposed AI checklist items were similar in both checklists and included the following: specifying of intended use and users; onsite and offsite requirements for generalisability assessment; inclusion and exclusion criteria at both participant and data input levels; version of AI system utilised; details on data acquisition; selection and pre-processing before analysis; handling of poor quality data; stating of the level of human-AI interaction and level of required expertise; specifying of the AI intervention and intended downstream outputs and their role in clinical pathways and clinical decision-making; detailing of the methods to identify errors and risk mitigation strategies; and the stating of access and license restrictions of the AI intervention. SPIRIT-AI recommended an additional item of describing pre-existing evidence regarding validation of the intended AI intervention (checklist item no. 6 in Table 1). Aside from researchers, clinicians could also use these checklist items during the clinical trial study appraisal processes to aid in evaluating the reproducibility, clinical translation and safety of the proposed AI intervention. A case example in which SPIRIT-AI or CONSORT-AI may be used is a prospective randomised controlled study of polyp detection on diagnostic colonoscopy using a real-time AI-assisted detection system that is compared with standard diagnostic colonoscopy without AI intervention.[36]

General guideline for early clinical evaluation of AI studies: DECIDE-AI

While SPIRIT-AI, CONSORT-AI, STARD-AI, QUADAS-AI, TRIPOD-AI and PROBAST-AI are specific to study design, DECIDE-AI is different and focuses on the early clinical evaluation stage and may be used across a variety of study designs (Table 2).[22] The DECIDE-AI guidelines included 17 AI-specific reporting items and 10 generic reporting items such as description of human factor tools, use cases considered, users involved, patient involvement and any significant change to the clinical workflow or care pathway caused by the AI system. Human factors, such as utility evaluation, safety and the effect of the intervention on the users' physical and cognitive performance, are important considerations in the regulatory process and acceptance of new AI interventions by patients, clinicians, regulatory bodies and potential investors.

In the inception of new AI interventions, developers of AI health technology should take note of available local guidelines that encompass regulatory requirements and legislation, which can assist in subsequent regulatory approval and commercial distribution within the local setting. For example, in Singapore, the Artificial Intelligence in Healthcare Guidelines[37] provide recommendations for both developers (early stage) and implementers of the AI technology (later stage) while taking into account specific legislation such as the Singapore Human Biomedical Research Act, and Personal Data and Protection Act.

Relevance of reporting guidelines at different phases of the AI innovation journey

After understanding the motivation and importance of each of the reporting consensus guidelines, one may then adopt the appropriate reporting guideline at applicable stages of the AI innovation (Fig. 1). First, the initial stage of the AI innovation journey is preclinical or algorithm development. A clinical need or an intended use is first identified and the research idea formulated and refined. Several factors should be taken into account when defining the specific intended use environment: the specific clinical problems, potential benefits, potential risk level, market size, patient demographics or ethnic groups, clinical settings (population-based vs clinic-based), hardware devices, users (patients vs healthcare professionals, general practitioners vs specialists), tasks (segmentation, classification, prediction), imaging specification (e.g. chest x-ray: posterior anterior vs anterior posterior vs lateral; computed tomography or magnetic resonance imaging: with or without contrast, axial vs sagittal vs coronal; retinal imaging: macula-centred vs optic disc-centred), and deployment mode (cloud-based, desktop-based or incorporation into edge devices). The definition of intended use environment is included in all the AI reporting guidelines, given its importance in regulations applicable to software as a medical device (SaMD).

Table 1. Comparison of SPIRIT-AI[21] and CONSORT-AI[20] checklists.

	SPIRIT 2013[21]	**SPIRIT-AI[21]**		**CONSORT 2010[20]**	**CONSORT-AI[20]**
Intended use	**Clinical trial protocol**	**AI-clinical trial protocol**		**Randomised controlled trial**	**AI-randomised controlled trial**
Checklist item no.			**Checklist item no.**		
1	Descriptive title identifying the study design, population, interventions and, if applicable, trial acronym	Indicate that the intervention involves artificial intelligence/ machine learning and specify the type of model Specify the intended use of the AI intervention	1	a) Identification as a randomised trial in the title b) Structured summary of trial design, methods, results, and conclusions	a) Indicate that the intervention involves artificial intelligence/ machine learning and specify the type of model b) State the intended use of the AI intervention within the trial in the title and/or abstract
2	Trial registration: a) Trial identifier and registry name. If not yet registered, name of intended registry. b) All items from the World Health Organization Trial Registration Dataset		23	Registration number and name of trial registry	
3	Protocol date and version identifier			-	
4	Funding: sources and types of financial, material and other support			-	
5	a) Names, affiliations and roles of protocol contributors b) Name and contact information for the trial sponsor c) Role of study sponsor and funders d) Composition, roles and responsibilities of the coordinating centre, steering committee, endpoint adjudication committee, data management team and other individuals/groups overseeing the trial			-	
6	a) Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention	a) Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users Describe any pre-existing evidence for the AI intervention.	2	a) Scientific background and explanation of rationale	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users
	b) Explanation for choice of comparators			-	
7	Specific objectives or hypotheses			b) Specific objectives or hypotheses	
8	Description of trial design, including type of trial, allocation ratio and framework		3	a) Description of trial design (such as parallel, factorial) including allocation ratio	
	-			b) Important changes to methods after trial commencement (such as eligibility criteria), with reasons	

Table 1. Comparison of SPIRIT-AI[21] and CONSORT-AI[20] checklists. (Cont'd)

	SPIRIT 2013[21]	SPIRIT-AI[21]		CONSORT 2010[20]	CONSORT-AI[20]
Intended use	**Clinical trial protocol**	**AI-clinical trial protocol**		**Randomised controlled trial**	**AI-randomised controlled trial**
Checklist item no.			**Checklist item no.**		
9	Description of study settings and list of countries where data will be collected Reference to where list of study sites can be obtained	Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting	4	b) Settings and locations where the data were collected	Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements
10	Inclusion and exclusion criteria for participants If applicable, eligibility criteria for study centres and individuals who will perform the interventions	State the inclusion and exclusion criteria at the (i) level of participants, AND at the (ii) level of input data		A) Eligibility criteria for participants	State the inclusion and exclusion criteria at the (i) level of participants, AND at the (ii) level of input data.
11	a) Interventions for each group with sufficient detail to allow replication, including how and when they will be administered	(i) State version of AI algorithm used (ii) Specify procedure for acquiring and selecting the input data for the AI intervention (iii) Specify the procedure for assessing and handling poor-quality or unavailable input data (iv) Specify whether there is human–AI interaction in the handling of the input data, and what level of expertise is required for users (v) Specify the output of the AI intervention (vi) Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice	5	The interventions for each group with sufficient details to allow replication, including how and when they were administered.	(i) State version of AI algorithm used (ii) Specify procedure for acquiring and selecting the input data for the AI intervention (iii) Specify the procedure for assessing and handling poor-quality or unavailable input data (iv) Specify whether there is human–AI interaction in the handling of the input data, and what level of expertise is required for users (v) Specify the output of the AI intervention (vi) Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice
	b) Criteria for discontinuing or modifying allocated interventions for a given trial participant			-	
	c) Strategies to improve adherence to intervention protocols and any procedures for monitoring adherence			-	
	d) Relevant concomitant care and intervention			-	

Table 1. Comparison of SPIRIT-AI[21] and CONSORT-AI[20] checklists. (Cont'd)

	SPIRIT 2013[21]	SPIRIT-AI[21]		CONSORT 2010[20]	CONSORT-AI[20]
Intended use	**Clinical trial protocol**	**AI-clinical trial protocol**		**Randomised controlled trial**	**AI-randomised controlled trial**
Checklist item no.			**Checklist item no.**		
12	Primary, secondary and other outcomes, including the specific measurement variable, analysis metric, method of aggregation, and time point for each outcome. Explanation of the clinical relevance of chosen efficacy and harm outcomes is strongly recommended		6	a) Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	
13	Time schedule of enrolment, interventions (including any run-ins and washouts), assessments and visits for participants. A schematic diagram is highly recommended			b) Any changes to trial outcomes after the trial commenced, with reasons	
14	Estimated number of participants needed to achieve study objectives and how it was determined, including clinical and statistical assumptions supporting any sample size calculations		7	a) How sample size was determined	
15	Strategies for achieving adequate participant enrolment to reach target sample size			b) When applicable, explanation of any interim analyses and stopping guidelines	
16	a) Method of generating the allocation sequence and list of any factors for stratification. To reduce predictability of a random sequence, details of any planned restriction should be provided in a separate document that is unavailable to those who enrol participants or assign interventions		8	a) Method used to generate the random allocation sequence b) Type of randomisation; details of any restriction (such as blocking and block size)	
	b) Mechanism of implementing the allocation sequence (e.g. central telephone; sequentially numbered, opaque, sealed envelopes), describing any steps to conceal the sequence until interventions are assigned		9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	
	c) Who will generate the allocation sequence, who will enrol participants, and who will assign participants to interventions		10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	
17	a) Who will be blinded after assignment to interventions and how		11	a) Blinding: If done, who was blinded after assignment to interventions (e.g. participants, care providers, those assessing outcomes) and how	
	b) If blinded, circumstances under which unblinding is permissible, and procedure for revealing a participant's allocated intervention during the trial			b) If relevant, description of the similarity of interventions	
18	a) Plans for assessment and collection of outcome, baseline and other trial data, including any related processes to promote data quality and a description of study instruments along with their reliability and validity, if known. Reference to where data collection forms can be found, if not in the protocol			-	
	b) Plans to promote participant retention and complete follow-up, including list of any outcome data to be collected for participants who discontinue or deviate from intervention protocols			-	
19	Plans for data entry, coding, security and storage, including any related processes to promote data quality. Reference to where details of data management procedures can be found, if not in the protocol			-	

Table 1. Comparison of SPIRIT-AI[21] and CONSORT-AI[20] checklists. (Cont'd)

	SPIRIT 2013[21]	**SPIRIT-AI[21]**		**CONSORT 2010[20]**	**CONSORT-AI[20]**
Intended use	**Clinical trial protocol**	**AI-clinical trial protocol**		**Randomised controlled trial**	**AI-randomised controlled trial**
Checklist item no.			**Checklist item no.**		
20	a) Statistical methods for analysing primary and secondary outcomes. Reference to where other details of the statistical analysis plan can be found, if not in the protocol		12	a) Statistical methods used to compare groups for primary and secondary outcomes	
	b) Methods for any additional analyses			b) Methods for additional analyses, such as subgroup analyses and adjusted analyses	
	c) Definition of analysis population relating to protocol non-adherence (e.g. as randomised analysis) and any statistical methods to handle missing data (e.g. multiple imputation)			-	
	-		13	a) For each group, the numbers of participants who were randomly assigned, received intended treatment and were analysed for the primary outcome	
	-			b) For each group, losses and exclusions after randomisation, together with reasons	
	-		14	a) Dates defining the periods of recruitment and follow-up	
	-			b) Why the trial ended or was stopped	
	-		15	A table showing baseline demographic and clinical characteristics for each group	
	-		16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	
	-		17	a) For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	
	-			b) For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
	-		18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	
21	a) Composition of data monitoring committee; summary of its role and reporting structure; statement of whether it is independent from the sponsor and competing interests; and reference to where further details about its charter can be found, if not in the protocol. Alternatively, an explanation of why a data monitoring committee is not needed			-	
	b) Description of any interim analyses and stopping guidelines, including who will have access to these interim results and make the final decision to terminate the trial			-	

Table 1. Comparison of SPIRIT-AI[21] and CONSORT-AI[20] checklists. (Cont'd)

	SPIRIT 2013[21]	SPIRIT-AI[21]		CONSORT 2010[20]	CONSORT-AI[20]
Intended use	**Clinical trial protocol**	**AI-clinical trial protocol**		**Randomised controlled trial**	**AI-randomised controlled trial**
Checklist item no.			**Checklist item no.**		
22	Plans for collecting, assessing, reporting, and managing solicited and spontaneously reported adverse events and other unintended effects of trial interventions or trial conduct	Specify any plans to identify and analyse performance errors. If there are no plans for this, justify why not	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, justify why not
23	Frequency and procedures for auditing trial conduct, if any, and whether the process will be independent from investigators and the sponsor			-	
24	Plans for seeking research ethics committee/ institutional review board approval			-	
25	Plans for communicating important protocol to relevant parties			-	
26	a) Who will obtain informed consent or assent from potential trial participants or authorised surrogates, and how (see item no. 32)			-	
	b) Additional consent provisions for collection and use of participant data and biological specimens in ancillary studies, if applicable			-	
27	How personal information about potential and enrolled participants will be collected, shared and maintained in order to protect confidentiality before, during and after the trial			-	
28	Financial and other competing interests for principal investigators for the overall trial and each study site			-	
			20	Trial limitations, addressing sources of potential bias, imprecision and, if relevant, multiplicity of analyses	
			21	Generalisability (external validity)	
			22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	
			24	Where the full trial protocol can be accessed, if available	
29	Statement of who will have access to the final trial dataset, and disclosure of contractual agreements that limit such access for investigators	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or reuse	25	Sources of funding and other support (such as supply of drugs), role of funders	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or reuse
30	Provisions, if any, for ancillary and post-trial care, and for compensation to those who suffer harm from trial participation			-	

Table 1. Comparison of SPIRIT-AI[21] and CONSORT-AI[20] checklists. (Cont'd)

	SPIRIT 2013[21]	**SPIRIT-AI[21]**		**CONSORT 2010[20]**	**CONSORT-AI[20]**
Intended use	**Clinical trial protocol**	**AI-clinical trial protocol**		**Randomised controlled trial**	**AI-randomised controlled trial**
Checklist item no.			**Checklist item no.**		
31	a) Plans for investigators and sponsor to communicate trial results to participants, healthcare professionals, the public and other relevant groups (e.g. via publication, reporting in results databases or other data sharing arrangements), including any publication restrictions			-	
	b) Authorship eligibility guidelines and any intended use of professional writers			-	
	c) Plans, if any, for granting public access to the full protocol, participant-level dataset and statistical code			-	
32	Model consent form and other related documentation given to participants and authorised surrogates			-	
33	Plans for collection, laboratory evaluation, and storage of biological specimens for genetic or molecular analysis in the current trial and for future use in ancillary studies, if applicable			-	

CONSORT: Consolidated Standards of Reporting Trials; CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence; SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence

Superscript numbers: Refer to REFERENCES

Second, researchers need to identify the appropriate datasets, which are known as the "dictionaries", to address the defined intended use environment. For data, it is always important to consider the following: data security (de-identification and extraction), labels (ground truth), size, type (structured vs unstructured data, cross sectional vs longitudinal, real-world vs clinical), phenotypes (positive vs control cases from different races), recruitment style with respect to inclusion and exclusion criteria (including the treatment of outliers), imaging devices, data heterogeneity (homogenous vs heterogenous), splitting of the training versus validation versus testing (local vs international), generalisability (representativeness of the target) and quality. In addition, the use of publicly available datasets may allow for reproducibility testing. At the initial research and development phase, the STARD-AI and QUADAS-AI guidelines could be used for reporting diagnostic accuracy and TRIPOD-AI and PROBAST-AI guidelines for the reporting of clinical prediction models using the retrospective or prospective datasets, although most AI researchers could utilise the retrospective datasets collected in the past decades to build the initial AI model.

Third, researchers need to select suitable technical methodologies (i.e. the "brain" that solves tasks) using machine learning or deep learning to analyse test datasets. For technical methodologies, it is important for AI researchers to design a robust AI architecture and operational flow based on the intended use environment. Depending on the data type, AI algorithms could be built as single modal versus multimodal (e.g. early vs late fusion) using machine learning (e.g. random forest, support vector machines, XGBoost) or deep learning (image-based, speech-based or natural language processing-based). For the overall AI architecture, it is critical to design an end-to-end system—from data input to pre-processing steps (contrast enhancement, image adjustment, cropping or centralisation); and inclusion of gradeability algorithm, appropriate machine learning or deep learning techniques, and explainability map using a selected visualisation technique (for image). It is always important to assess the robustness of AI algorithms using a pre-set operating threshold (on the training/validation datasets) to evaluate it on the local or international testing datasets using performance metrics such as the area under the receiver operating characteristic curve, sensitivity, specificity and 95% confidence interval.

Upon completion of AI algorithm and development, researchers should explore the potential technical disclosure of the AI algorithm with the relevant

Table 2. Comparing SPIRIT-AI, CONSORT-AI and DECIDE-AI.

	SPIRIT-AI	CONSORT-AI	DECIDE-AI
Extension of existing guideline	Yes	Yes	No
Stage of intended use	Late clinical trial (phase III) Comparative prospective evaluation	Late clinical trial (phase III) Comparative prospective evaluation	Early clinical trial (phase I and II)
Study design	Clinical trial protocol	Randomised controlled trial	Any
Key focus	Standardisation of reporting	Standardisation of reporting	Assessing clinical utility, safety and human factors

CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Drive by Artificial Intelligence; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence

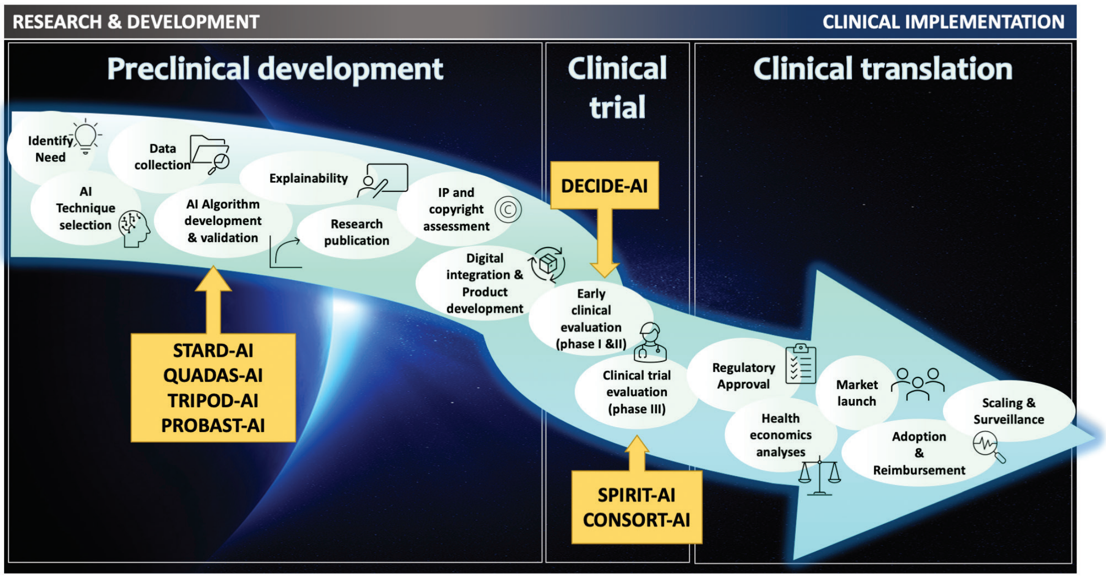


Fig. 1. Stages of the AI innovation.
AI: artificial intelligence; CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Drive by Artificial Intelligence; IP: intellectual property; PROBAST-AI: Prediction Model Risk of Bias Assessment Tool–Artificial Intelligence; QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies–Artificial Intelligence; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence; STARD-AI: Standards for Reporting of Diagnostic Accuracy Study–Artificial Intelligence; TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis or Diagnosis–Artificial Intelligence

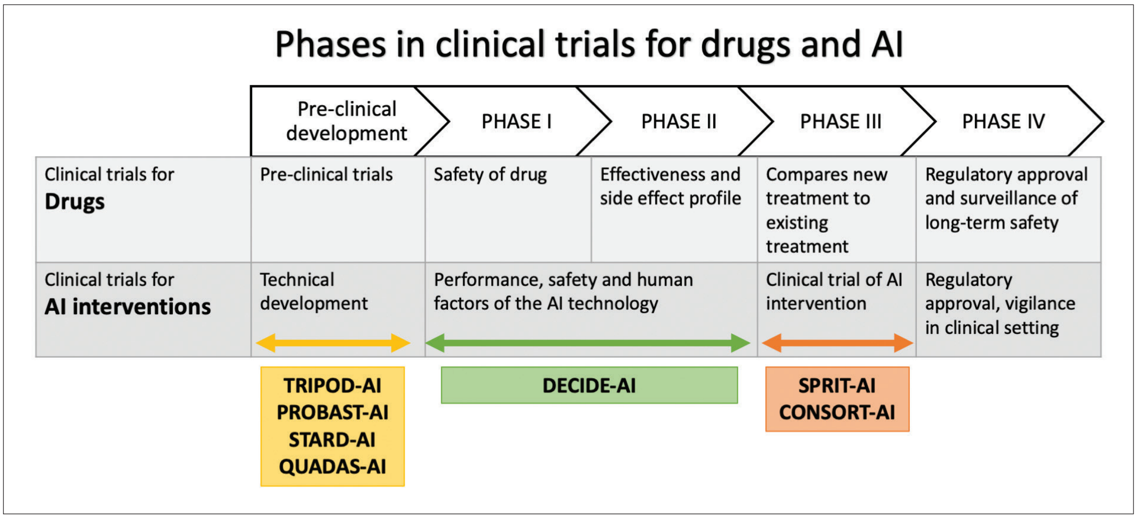
institutional intellectual property (IP) office should the findings be robust. Given that many AI algorithms are developed using off-the-shelf technical packages, the technology transfer office and IP office often find it challenging to file patents that rely heavily on AI systems for discovery, although these algorithms may be kept as know-hows, trade secrets and intellectual/commercial property. It is always important to address IP-related issues prior to any publication of AI algorithms in the scientific domain to avoid invalidating any potential patent.

Fourth, once the AI technology is deemed to be robust and mature, it will enter the clinical trial stage (Fig. 1). Ideally, a user-friendly and graphical user interface and user experience for the AI algorithms should be deployed at the intended settings when the clinical trials are conducted prospectively. In the early-stage clinical evaluation of AI-based decision support system, the DECIDE-AI guideline could be used to evaluate the performance, safety and human factors of the AI technology, a phase which is equivalent to phase I and II of pharmaceutical trials (Fig. 2), followed by a larger-scale clinical evaluation using either SPIRIT-AI or CONSORT-AI for clinical trial protocol reporting or randomised controlled trial reporting, respectively. A comparison of the original SPIRIT and CONSORT guidelines with their AI extensions can be found in Table 1. Similar to designing conventional clinical trials, researchers would need to determine the AI-based trial design: randomised controlled versus non-randomised; allocation ratios; inclusion versus exclusion criteria; superiority versus non-inferiority trials; sample size; recruitment sites; and determination of the internationally acceptable gold standards for comparative trials. In addition, in the context of interventional studies, novel trial designs that harness AI-based platforms can potentially identify more responders to treatment.[38,39]

Finally, upon successful completion of clinical trials, the AI technology then moves into the last stage: clinical translation and real-world development (Figs. 1 and 3). Regulatory approval should then be obtained. Regulatory guidelines, such as the International Medical Device Regulators Forum SaMD guidelines used by the US Food and Drug Administration,[40] and local guidelines, such as the Health Sciences Authority regulatory guidelines for SaMD[41] and the Artificial Intelligence in Healthcare Guidelines,[37] may be used.

Beyond regulatory approval, implementation research, health services research and workforce training and education are key aspects to facilitate clinical adoption and implementation (Fig. 3). Health economic analyses including cost-utility analysis, cost-effective analysis, cost-minimisation analysis and cost-benefit analysis may also be performed to assist in acceptance of the AI technology at a policy level.[42] This information can help clinicians, patients and policymakers evaluate the potential applicability and impact of such AI technologies in real-world clinical settings. Timely engagement with key stakeholders including patients, clinicians, healthcare delivery organisation leaders, operational personnel, policymakers, researchers, funders, product manufacturers and relevant medical societies is essential for convergence science, operational changes and actual clinical implementation. Once the AI technology is widely adopted, it is important to evaluate its impact on the population or global health and the society (Fig. 3).

In summary, AI reporting guidelines serve as useful guides for AI developers or users to build and appraise different AI technologies in health at different stages of innovation. To build a robust and clinically useful AI algorithm, it is important to define the intended use; choose or build the right datasets, technical methodology and architecture using the different reporting guidelines; and evaluate the performance using the appropriate statistical analyses. Given the prevailing and evolving SaMD rules, early engagement with applicable



Phases in clinical trials for drugs and AI

	Pre-clinical development	PHASE I	PHASE II	PHASE III	PHASE IV
Clinical trials for **Drugs**	Pre-clinical trials	Safety of drug	Effectiveness and side effect profile	Compares new treatment to existing treatment	Regulatory approval and surveillance of long-term safety
Clinical trials for **AI interventions**	Technical development	Performance, safety and human factors of the AI technology		Clinical trial of AI intervention	Regulatory approval, vigilance in clinical setting
	TRIPOD-AI **PROBAST-AI** **STARD-AI** **QUADAS-AI**	**DECIDE-AI**		**SPRIT-AI** **CONSORT-AI**	

Fig. 2. Phases in clinical trials for drugs and AI.

AI: artificial intelligence; CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence; DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Drive by Artificial Intelligence; PROBAST-AI: Prediction Model Risk of Bias Assessment Tool–Artificial Intelligence; QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies–Artificial Intelligence; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence; STARD-AI: Standards for Reporting of Diagnostic Accuracy Study–Artificial Intelligence; TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis or Diagnosis–Artificial Intelligence

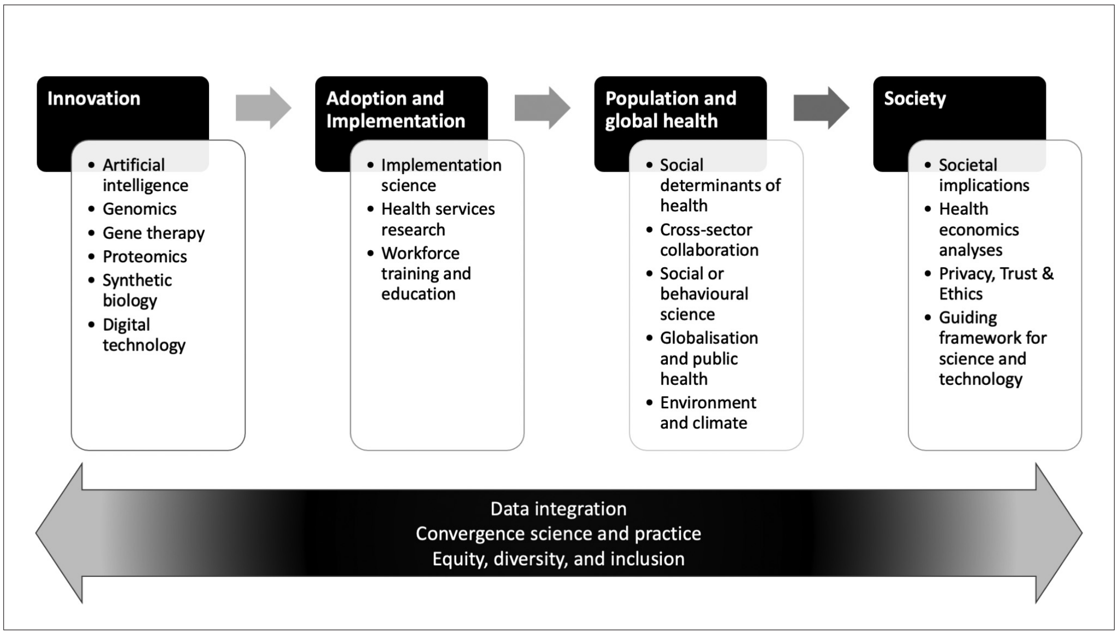

Innovation
• Artificial intelligence
• Genomics
• Gene therapy
• Proteomics
• Synthetic biology
• Digital technology
Adoption and Implementation
• Implementation science
• Health services research
• Workforce training and education
Population and global health
• Social determinants of health
• Cross-sector collaboration
• Social or behavioural science
• Globalisation and public health
• Environment and climate
Society
• Societal implications
• Health economics analyses
• Privacy, Trust & Ethics
• Guiding framework for science and technology
Data integration
Convergence science and practice
Equity, diversity, and inclusion

Fig. 3. Key aspects to facilitate clinical adoption, implementation and evaluation of AI innovation.

technology transfer or IP offices and regulatory bodies is key. At the same time, rapid developments in AI research mean that the field must be prepared to adapt and evolve as new techniques emerge. AI has enormous potential to enhance clinical outcomes and experiences for patients. These reporting guidelines are excellent initiatives that involve multiple stakeholders originating from diverse backgrounds (clinicians, allied health professionals, technology developers, bioethicists, patient representatives, industry, regulatory bodies, government and journal editorial boards), and they could play a pivotal role not only in the research settings but also in clinical and governmental policymaking settings to ensure that the continuum of AI innovation from robust and rigorous evaluation to clinical deployment and reimbursement reaches fruition for all AI technologies in medicine.

Disclosure

Dr Daniel SW Ting holds a patent on a deep learning system for detection of retinal diseases, co-founded and holds equity of EyRIS Singapore. Dr Carolyn SP Lam holds a patent on a deep learning system for detection of cardiac disease, co-founded and holds equity in Us2.ai. Dr Dean Ho is scientific co-founder and shareholder of KYAN Therapeutics. He is also a co-inventor of pending patents pertaining to AI-based drug development and personalised medicine.

REFERENCES

1. Ting DSW, Carin L, Dzau V, et al. Digital technology and COVID-19. Nat Med 2020;26:459-61.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
3. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24-9.
4. Ting DSW, Liu Y, Burlina P, et al. AI for medical imaging goes deep. Nat Med 2018;24:539-40.
5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
6. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211-23.
7. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284:574-82.
8. Cleverley J, Piper J, Jones MM. The role of chest radiography in confirming covid-19 pneumonia. BMJ 2020;370:m2426.
9. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18.
10. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583-9.
11. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596:590-6.
12. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020; 577:706-10.
13. Gatenby RA, Silva AS, Gillies RJ, et al. Adaptive therapy. Cancer Res 2009;69:4894-903.

14. Ng WY, Tan TE, Movva PVH, et al. Blockchain applications in health care for COVID-19 and beyond: A systematic review. Lancet Digit Health 2021;3:e819-29.

15. Tan TE, Anees A, Chen C, et al. Retinal photograph-based deep learning algorithms for myopia and a blockchain platform to facilitate artificial intelligence medical research: A retrospective multicohort study. Lancet Digit Health 2021;3:e317-29.

16. Ng WY, Tan TE, Xiao Z, et al. Blockchain technology for ophthalmology: Coming of age? Asia Pac J Ophthalmol (Phila) 2021;10:343-7.

17. Chen RJ, Lu MY, Chen TY, et al. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng 2021;5:493-7.

18. Tosun AB, Pullara F, Becich MJ, et al. Explainable AI (xAI) for anatomic pathology. Adv Anat Pathol 2020;27:241-50.

19. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: An economic analysis modelling study. Lancet Digit Health 2020;2:e240-9.

20. Liu X, Cruz Rivera S, Moher D, et al; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Lancet Digit Health 2020;2:e537-48.

21. Rivera SC, Liu X, Chan AW, et al; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. BMJ 2020;370:m3210.

22. Vasey B, Nagendran M, Campbell B, et al; DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28:924-33.

23. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021;27:1663-5.

24. Sounderajah V, Ashrafian H, Golub RM, et al; STARD-AI Steering Committee. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol. BMJ Open 2021;11:e047709.

25. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digit Health 2019;1:e271-97.

26. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689.

27. Chesterman S. We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law. Cambridge, UK: Cambridge University Press; 2021.

28. Blasiak A, Truong A, LWJ Tan, et al. PRECISE CURATE.AI: A prospective feasibility trial to dynamically modulate personalized chemotherapy dose with artificial intelligence. J Clin Oncol 2022;40(16 Suppl):1574.

29. Pantuck AJ, Lee DK, Kee T, et al. Artificial intelligence: Modulating BET bromodomain inhibitor ZEN-3694 and enzalutamide combination dosing in a metastatic prostate cancer patient using CURATE.AI, an artificial intelligence platform. Adv Therap 2018;1:1800104.

30. Taylor M, Liu X, Denniston A, et al; SPIRIT-AI and CONSORT-AI Working Group. Raising the bar for randomized trials involving artificial intelligence: The SPIRIT-Artificial Intelligence and CONSORT-Artificial Intelligence guidelines. J Invest Dermatol 2021;141:2109-2111.

31. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11:e048008.

32. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ 2015;350:g7594.

33. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann Intern Med 2019;170:W1-33.

34. Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 Statement: Defining standard protocol items for clinical trials. Rev Panam Salud Publica 2015;38:506-14.

35. Moher D, Hopewell S, Schulz KF, et al; CONSORT. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. Int J Surg 2012;10:28-55.

36. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. Gut 2019;68(10):1813-9.

37. Health Sciences Authority. Artificial intelligence in healthcare guidelines (AIHGIe). October 2021. https://www.moh.gov.sg/docs/librariesprovider5/eguides/1-0-artificial-in-healthcare-guidelines-(aihgle)_publishedoct21.pdf. Accessed 19 April 2023.

38. Tan BKJ, Teo CB, Tadeo X, et al. Personalised, Rational, Efficacy-Driven Cancer Drug Dosing via an Artificial Intelligence SystEm (PRECISE): A protocol for the PRECISE CURATE.AI pilot clinical trial. Front Digit Health 2021;3:635524.

39. Blasiak A, Kee TW, Rashid MBM, et al. CURATE.AI-optimized modulation for multiple myeloma: An N-of-1 randomized trial. Cancer Res 2020;80(16 Suppl):CT268.

40. International Medical Device Regulators Forum. Software as a medical device (SaMD): Clinical evaluation. 21 September 2017. https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation. Accessed 19 April 2023.

41. Health Sciences Authority. Regulatory guidelines for software medical devices – A life cycle approach. April 2022. https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance-documents-for-medical-devices/regulatory-guidelines-for-software-medical-devices---a-life-cycle-approach_r2-(2022-apr)-pub.pdf. Accessed 19 April 2023.

42. Kwee A, Teo ZL, Ting DSW. Digital health in medicine: Important considerations in evaluating health economic analysis. Lancet Reg Health West Pac 2022;23:100476.