BMJ Health & Care Informatics

Evaluation framework to guide implementation of AI systems into healthcare settings

Sandeep Reddy ,[1] Wendy Rogers,[2] Ville-Petteri Makinen,[3] Enrico Coiera ,[4] Pieta Brown,[5] Markus Wenzel,[6] Eva Weicken,[6] Saba Ansari,[7] Piyush Mathur ,[8] Aaron Casey,[3] Blair Kelly[7]

To cite: Reddy S, Rogers W, Makinen V-P, *et al*. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform* 2021;**28**:e100444. doi:10.1136/bmjhci-2021-100444

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100444).

Received 08 July 2021
Accepted 30 September 2021

For numbered affiliations see end of article.

Correspondence to
Dr Sandeep Reddy;
sandeep.reddy@deakin.edu.au

ABSTRACT

Objectives To date, many artificial intelligence (AI) systems have been developed in healthcare, but adoption has been limited. This may be due to inappropriate or incomplete evaluation and a lack of internationally recognised AI standards on evaluation. To have confidence in the generalisability of AI systems in healthcare and to enable their integration into workflows, there is a need for a practical yet comprehensive instrument to assess the translational aspects of the available AI systems. Currently available evaluation frameworks for AI in healthcare focus on the reporting and regulatory aspects but have little guidance regarding assessment of the translational aspects of the AI systems like the functional, utility and ethical components.

Methods To address this gap and create a framework that assesses real-world systems, an international team has developed a translationally focused evaluation framework termed 'Translational Evaluation of Healthcare AI (TEHAI)'. A critical review of literature assessed existing evaluation and reporting frameworks and gaps. Next, using health technology evaluation and translational principles, reporting components were identified for consideration. These were independently reviewed for consensus inclusion in a final framework by an international panel of eight expert.

Results TEHAI includes three main components: capability, utility and adoption. The emphasis on translational and ethical features of the model development and deployment distinguishes TEHAI from other evaluation instruments. In specific, the evaluation components can be applied at any stage of the development and deployment of the AI system.

Discussion One major limitation of existing reporting or evaluation frameworks is their narrow focus. TEHAI, because of its strong foundation in translation research models and an emphasis on safety, translational value and generalisability, not only has a theoretical basis but also practical application to assessing real-world systems.

Conclusion The translational research theoretic approach used to develop TEHAI should see it having application not just for evaluation of clinical AI in research settings, but more broadly to guide evaluation of working clinical systems.

INTRODUCTION

Progress in artificial intelligence (AI) has opened new opportunities to respond to many healthcare-related issues.[1] However, recent AI systems have fallen short of their translational goals.[2–4] AI systems are often developed from a technical perspective, with consideration of how they fit into and value to real-world workflows a secondary concern. Machine learning may be applied to biased or poor-quality data sets[5] Further, technology-supported clinical decisions require a robust ethical framing which is not considered in purely technical evaluations.[4] Using and integrating AI systems in clinical settings can be potentially expensive and disruptive, thus necessitating strong justification for their deployment.[2 3] As a result of, it is sometimes difficult to have confidence in the generalisability of the AI systems and adopters may face unnecessary roadblocks on the path to an effective healthcare response.[3 4] Therefore, a rigorous evaluation that assesses AI systems early and at various stages of their clinical deployment, is crucial.[2 4]

Currently available evaluation frameworks for AI systems in healthcare generally focus on reporting and regulatory aspects.[6–8] This is helpful when you have AI systems deployed in healthcare services and integrated with clinical workflow. However, despite numerous such evaluation and reporting frameworks, it is evident there is an absence of an evaluation framework that assesses various stages of development, deployment, integration and adoption of AI systems. Dependance on disparate evaluation frameworks to assess different aspects and phases of AI systems is unrealistic. Also, currently available evaluation and reporting frameworks fall short in adequately assessing the functional, utility and ethical aspects of the models despite growing evidence about the limited

adaptability of AI systems in healthcare. The absence especially of an assessment of the ethical dimensions such as privacy, non-maleficence and explainability in the available frameworks indicates their inadequacy in providing an inclusive and translational evaluation. Therefore, a comprehensive yet practical instrument that assess the translational aspects and various phases of available AI systems is required.

METHODOLOGY

The Declaration of Innsbruck describes evaluation of information and communication technology as 'the act of measuring or exploring properties of a health information system (in planning, development, implementation or operation), the result of which informs a decision to be made concerning that system in a specific context.'[9] An assessment framework adopting the declaration thus must consider that the health information system not only includes the software and hardware but also the environment encompassing the actors and their interactions.[10] However, evaluation should be limited to assessing the specific evidence required to make a given decisions; if not, evidence gathering may become wasteful or infeasible.

This approach aligns with principles for translational research (TR) which focuses on facilitating the translation of scientific evidence into real-world practice[11] and with Health Technology Assessment, which is the systematic evaluation of health technologies and interventions.[12] Translation research principles support processes that turn observations in the laboratory or clinic or community into interventions that improve the health of individuals, encourage multidisciplinary collaboration and enables adoption of evidence-based approaches. These principles have guided the development of the Translational Evaluation of Healthcare AI (TEHAI) framework.

As per this approach and as a first step, we adopted a critical review of related literature including frameworks and guidelines (covering AI in healthcare reporting and evaluation).[6–8 13–15] The project team identified key components that could be considered in the framework and developed a draft initial version of the 'TEHAI' framework. Candidate components and subcomponents were identified using a consensual approach that explored their validity and relevance to the development of AI systems in healthcare. The draft framework was then reviewed by an eight-member international panel with expertise in medicine, data science, healthcare policy, biomedical research and healthcare commissioning drawn from the UK, USA and New Zealand. Panel members were provided the framework and documentation relating to the framework including a guide to interpret each component and subcomponent. Feedback from the expert panel was used to refine and craft the final version of TEHAI.

RESULTS

TEHAI includes three main components capability, adoption and utility to assess various AI systems. The components are a synthesis of several activities and were chosen with a focus on the translational aspects of AI in healthcare, that is, how is AI applied and used in healthcare? The emphasis on translational and ethical features of the model development and deployment distinguishes TEHAI from other evaluation instruments. Outlined in figure 1 and the subsequent narrative are high level details of TEHAI's 3 main components and its 15 subcomponents. The description of components is kept brief to facilitate their use within a checklist. Full details of the framework including the scoring system are outlined in online supplemental file 1.

Capability

This component assesses the intrinsic technical capability of the AI system to perform its expected purpose, by reviewing key aspects as to how the AI system was developed. Unless the model has been trained and tested appropriately, it is unlikely the system will be useful in healthcare environments.

Objective

This subcomponent assesses whether the system has an ethically justifiable objective, that is, stated contribution to addressing an identified healthcare problem with the aims of reducing morbidity and/or mortality and/or increasing efficiency. This subcomponent is scored on a scale of how well the objective is articulated, that is, the problem the AI addresses, why the study is being conducted and how it adds to the body of knowledge in the domain are clearly articulated.

Dataset source and integrity

An AI system is only as good as the data it was derived from.[13] If the data do not reflect the intended purpose, the model predictions are likely to be useless or even harmful.[4] This subcomponent evaluates the source of the data and the integrity of datasets used for training and testing the AI system including an appraisal of the representation and coverage of the target population in the data, and the consistency and reproducibility of the data collection process. Scoring is determined by how well the dataset is described, how well the dataset fits with the objective from a technical point of view and how credible/reliable the data source is. This subcomponent also considers when new data are acquired to train a clinically embedded model that appropriate checks are undertaken to ensure integrity and alignment of data to previously used data.

Internal validity

An internally valid model will be able to predict health outcomes reliably and accurately within a predefined set of data resources that were used wholly or partially when training the model.[16] This validation process includes the classical concept of goodness-of-fit, but also

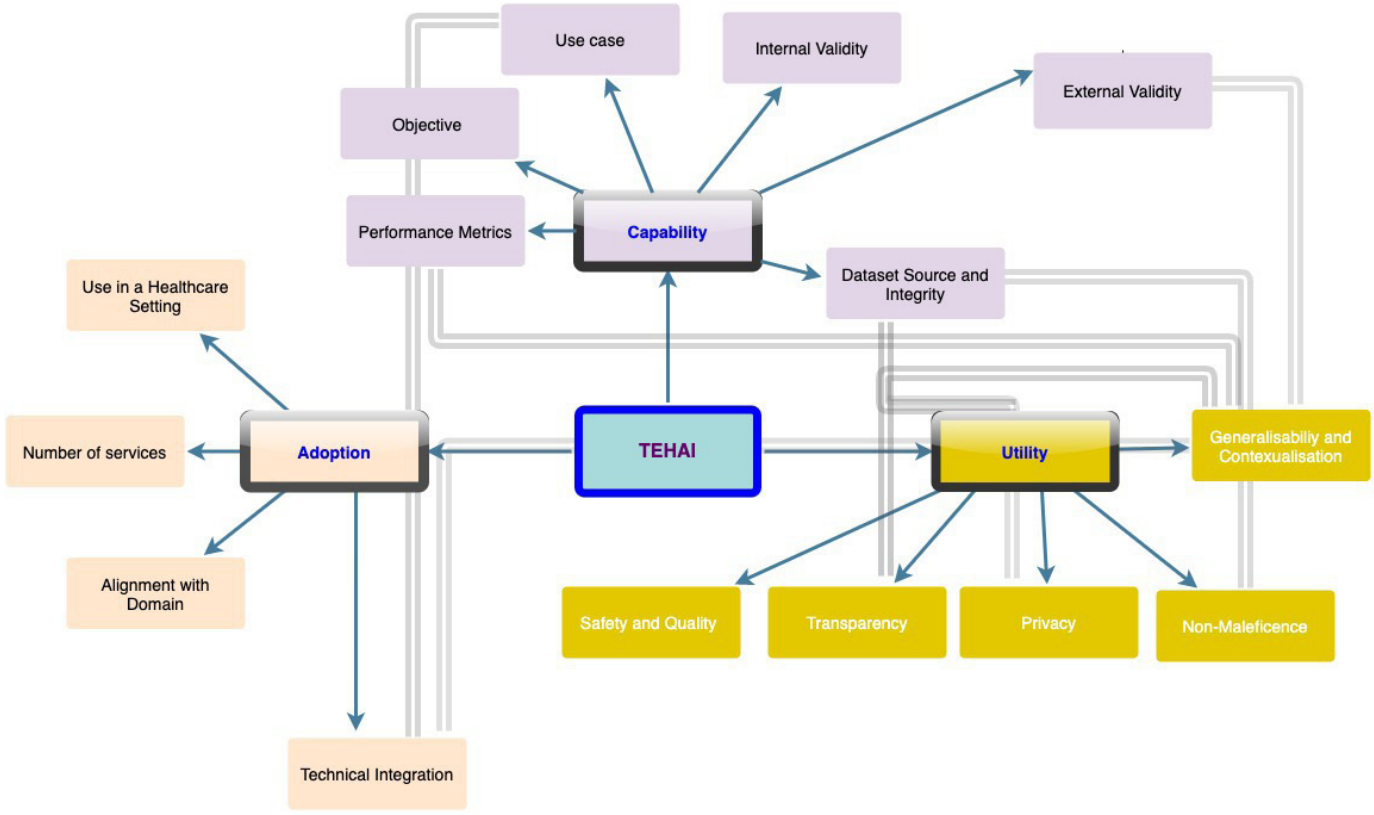

Figure 1 Translational evaluation of healthcare AI. Three main components—capability, utility and adoption, with 15 subcomponents. The components and associated subcomponents are represented in the same colour. Subcomponents with cross-relationships are linked by bold arrows. AI, artificial intelligence; TEHAI, Translational Evaluation of Healthcare AI.

cross-validation schemes that derive training and test sets from the same sources of data. Scoring is based on the size and properties of the training data set with respect to the healthcare challenge, the diversity of the data to ensure good modelling coverage, and whether the statistical performance of the model (eg, in a classification task) is high enough to satisfy the requirements of usefulness in the healthcare context.

External validity

To evaluate external validity, we investigate whether the external data used to assess model performance came from substantially distinct external sources that did not contribute any data towards model training.[17] Examples of external data sources include independent hospitals, institutions or research groups that were not part of the model construction team or a substantial temporal difference between the training and validation data collections. The scoring is based on the size and coverage of the external data (if any) and whether there is sufficient variation in the external data to allow meaningful statistical conclusions.

Performance metrics

Performance metrics refer to mathematical formulas that are used for assessing how well an AI model predicts clinical or other health outcomes from the data.[5 16 17] These performance metrics can be classification or regression or qualitative metrics. If the metrics are chosen poorly, it is not possible to assess the accuracy of the models reliably. Furthermore, specific metrics have biases, which means that a combination of multiple metrics might lead to more reliable conclusions in some cases. This subcomponent examines whether appropriate performance measures relevant to the given task had been selected for the presentation of the study results. Metrics are also evaluated for their reliability across domains or when models are updated with new evidence if such iterative tool development is a likely scenario. This subcomponent is scored according to how well the performance metrics fit the study and how reliable they are likely to be considering the nature of the healthcare challenge.

Use case

This subcomponent investigates the justification for the use of AI for the study as opposed to assessing statistical or analytical methods. This tests if the study has considered the relevance and fit of the AI to the particular healthcare domain it is being applied to. This subcomponent is scored on a scale of how well the use case is stated that is, whether the study presents evidence or arguments to justify the AI method used.

Utility

This component evaluates the usability of the AI system across different dimensions including the contextual

relevance, and safety and ethical considerations regarding eventual deployment into clinical practice. It also assesses the efficiency of the system (achieving maximum productivity while working in a competent manner) as evaluated through the quality, adoption and alignment measures. Utility as measured through these dimensions assesses the applicability of the AI system for the particular use case and the domain in general.

Generalisability and contextualisation

Biases or exacerbation of disparities due to underrepresentation or inappropriate representation within datasets used both in training and validation can have an adverse and potentially unjust effect on the real-world utility of an AI model. This subcomponent is scored based on how well an AI model is expected to capture the specific groups of people it is most intended for. Scoring also considers contextualisation. The context of an AI application is defined here as the alignment between the model's performance, expected results, characteristics of the training data and the overall objective.

Safety and quality

It is critical that AI models being deployed in healthcare, especially in clinical environments, are assessed for their safety and quality.[13 18] Appropriate consideration should be paid to the presence of ongoing monitoring mechanisms in the study, such as adequate clinical governance that will provide a systematic approach to maintaining and improving the safety and quality of care within a healthcare setting. This subcomponent is scored based on the presence and strength of any safety and quality evaluations and how likely they are to ensure safety and quality when AI is applied in the real world.

Transparency

This subcomponent assesses the extent to which model functionality and architecture is described in the study and the extent to which decisions reached by the algorithm are understandable (ie, black box or interpretable). Relevant elements include the overall model structure, the individual model components, the learning algorithm and how the specific solution is reached by the algorithm. This subcomponent is scored on a scale of how transparent, interpretable and reproducible the AI model is given the information available.

Privacy

This subcomponent refers to personal privacy, data protection and security. This subcomponent is ethically relevant to the concept of autonomy/self-determination, including the right to control access to and use of personal information, and the consent processes used to authorise data uses. Privacy is scored on the extent to which privacy considerations are documented, including consent by study subjects, the strength of data security and data life cycle throughout the study itself and consideration for future protection if deployed in the real world.

Non-maleficence

This subcomponent refers to the identification of actual and potential harms, beyond patient safety, caused by the AI and any actions taken to avoid foreseeable or unintentional harms. Harms to individuals may be physical, psychological, emotional or economic.[13 18] Harms may affect systems/organisations, infrastructure and social well-being. This subcomponent is scored on the extent to which potential harms of the AI are identified, quantified and the measures taken to avoid harms and reduce risk.

Adoption

There have been issues with the adoption and integration of AI systems in healthcare delivery even with those that have demonstrated their efficacy, although in in-silico or controlled environments. Therefore, it is important to assess the translational value of current AI systems. This component appraises this by evaluating key elements that demonstrate the adoption of the model in real life settings.

Use in a healthcare setting

As many AI systems have been developed in controlled environments or in silico there is a need to assess for evidence of use in real world environments and integration of new AI models with existing health service information systems. Also, the trials may have demonstrated efficacy, but a 'real-world' deployment is necessary to demonstrate effectiveness. It is important to consider the utility of the system for its users and its beneficiaries, for example, users might be clinicians and administrators, while beneficiaries might be patients. Both elements reflect the sustainability of the system in the service and its acceptance by patients and clinicians. This subcomponent is scored according to the extent to which the model has been integrated into external healthcare sites and the utility of the system for end users and beneficiaries.

Technical integration

This subcomponent evaluates how well the models integrate with existing clinical/administrative workflows outside of the development setting, and their performance in such situations. In addition, the subcomponent includes reporting of failed integration where it occurs, that is, even if the model performs poorly it is reported. This subcomponent is scored according to how well the integration aspects of the model are anticipated and if specific steps to facilitate practical integration have been taken.

Number of services

Many AI in healthcare studies are based on single site use without evidence of wider testing or validation. In this subcomponent, we review reporting quantitative assessment of wider use. This subcomponent is scored according to how well the use of the system across multiple healthcare organisations and/or multiple types of healthcare environments is described.

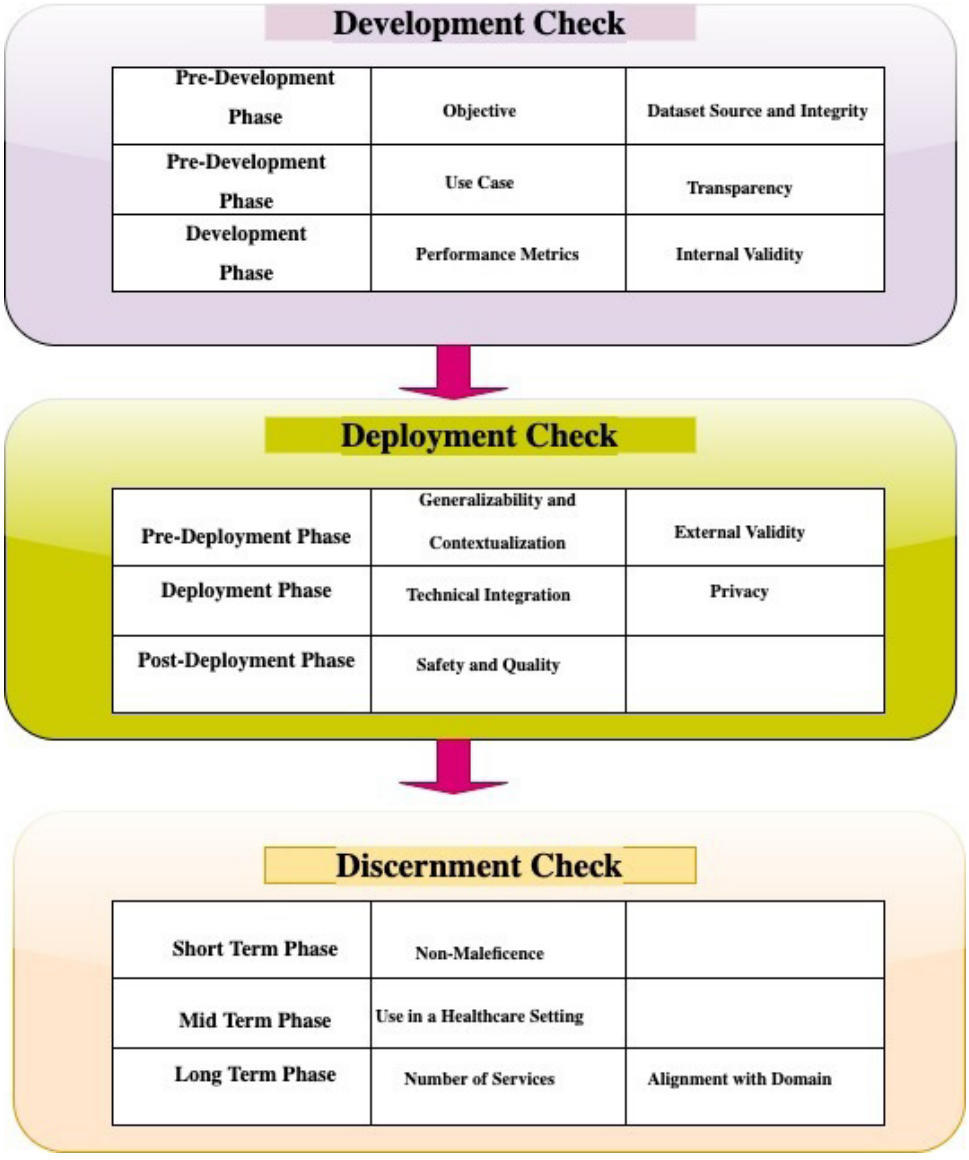

Figure 2 TEHAI checks during different phases. Three main phases including development, deployment and discernment phases with various subcomponents. TEHAI, Translational Evaluation of Healthcare Artificial Intelligence.

Alignment with domain

This category considers the alignment and relevance of the AI system to the particular healthcare domain and its likely long-term acceptance. In other words, the model is assessing the benefits of the AI system to the particular medical domain the model is being applied to. This again relates to the translational aspects of the AI model. This subcomponent is scored according to how well the benefits of the AI system for the medical domain are articulated.

To help with the implementation of these checklist, we recommend below the different phases (figure 2) when the subcomponents have to be checked. The checks can be performed when the AI system is being developed (Development Check), when the AI system is being deployed (Deployment Check), and as part of ongoing monitoring (Discernment Check).

Scoring

The scoring of evidence of the various subcomponents is organised in a matrix (figure 3). The initial scoring per subcomponent is based on a scale of 0–3 based on the degree to which the criteria in the subcomponents are met, that is, presence and absence of features in the AI system (see the additional information section for the details of what is being scored and how). Then, the awarded score is multiplied by a weight allocated to the subcomponent. The weighting process in scoring indicates certain criteria are more important than others and will, when the weights are combined with the scores, provide a more distinguishable overall score. We have allocated weights to each subcomponent in the framework to recognise the importance of certain subcomponents to translational medicine and provide more granularity to the process. The weights are either 5 or 10 indicating

Component	Sub-component	Initial Score	Weight	Subcomponent Score= Initial Score x Weight	
Capability	*Objective of Study*	0-3	10	Weight 5	0-9 10-14 15 and above
	Dataset Source and Integrity		10		
	Internal Validity		10		
	External Validity		10		
	Performance Metrics		10		
	Use Case		5		
Utility	*Generalizability and Contextualisation*	0-3	10	Weight 10	0-19 20-29 30 and above
	Safety and Quality		10		
	Transparency		10		
	Privacy		10		
	Non-Maleficence		10		
Adoption	*Use in a Healthcare Setting*	0-3	10		
	Technical Integration		10		
	Number of Services		5		
	Alignment with Domain		5		

Figure 3 TEHAI scoring matrix. Scores and weights for the assessment of each subcomponent of capability, utility and adoption. TEHAI, Translational Evaluation of Healthcare Artificial Intelligence.

midpoint and endpoint on a scale of 10. Each subcomponent is allocated a specific weight based on the team's view of their degree of importance to the evaluation.

The final subcomponent scores, which are the multiplied values of the score and weight will be highlighted in a traffic lights colour scheme. Each weight has its own traffic light scaling system to ensure the equivalency of final scores immaterial of what weight is employed. There will be no overall component or evaluation score as this approach will potentially obscure individual subcomponent strengths or weaknesses and mislead readers. Of note, evaluators may mix and match different subcomponents at various stages of development and deployment of the system (figure 2). While the scoring system may seem complex, when set up in a spreadsheet or a database can be easily automated to minimise the need for manual calculation.

DISCUSSION

The application of AI in healthcare, driven by recent advances in machine learning, is growing and will likely continue to do so.[19] Such application requires appropriate datasets among other key resources.[1] Obtaining such datasets can prove difficult, meaning many AI developers rely on whatever is available to them to produce initial results.[13] In certain instances, comprehensive evaluation of AI may not occur until the model is deployed due to limited internal evaluation capacity or an excessive focus on predeployment evaluation.[2] Awaiting evaluation after the model is deployed in clinical practice presents a safety and quality risk.[4] Therefore, evaluating AI models predeployment and postdeployment along the AI-life cycle can identify potential concerns and issues with the model, avoiding harmful effects on patients' outcomes and clinical decision making.

One of the major limitations with many existing reporting or evaluation frameworks is their narrow focus. Some focus on reporting of clinical trials evaluating AI interventions[6 7] on a specific medical domain[20 21] or compare a particular type of AI model to human clinicians[8] limiting the generalisability of such frameworks. It is now increasingly becoming evident that many AI systems, that have shown promise in in-silico environment or when deployed in single sites, are not fit for purpose when deployed widely.[3 5] Therefore, evaluation of AI systems not only has to commence earlier in the development process but also must be continuous and comprehensive, which is lacking in many currently available evaluation and reporting frameworks.

TEHAI, because of its strong foundation in TR and an emphasis on safety, translational value and generalisability, has not only a vigorous theoretical basis but also practical appeal in that it is designed to assess real-world systems. As not all developers or health services will have the resources to use TEHAI in its entirety, it offers some flexibility by demarcating the three components of capability, utility and adoption, each of which is independently scored. TEHAI is designed to be used at various phases of AI model development, deployment and workflow integration in addition to considering the translational and ethical aspects of the AI model in question, thereby providing a more comprehensive yet flexible assessment framework.

CONCLUSION

The monitoring and evaluation of AI in healthcare should be de rigueur and requires an appropriate evaluation framework for regulatory agencies and other bodies. Health services may also need to evaluate AI applications for safety, quality and efficacy, before their adoption and integration. Further, developers and vendors may want to assess their products before regulatory approval and release into the market. TEHAI, because of its comprehensiveness and flexibility to different stages of AI development and deployment, may be of use to all these groups.

Author affiliations
[1]School of Medicine, Deakin University, Geelong, Victoria, Australia
[2]Department of Philosophy, Macquarie University, Sydney, New South Wales, Australia
[3]South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia
[4]Australian Institute of Health Innovation, Macquarie University, Sydney, New South Wales, Australia
[5]Orion Health, Auckland, Auckland, New Zealand
[6]Fraunhofer Institute for Telecommunications Heinrich-Hertz-Institute HHI, Berlin, Germany
[7]Deakin University Faculty of Health, Geelong, Victoria, Australia
[8]Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio, USA

Acknowledgements The authors wish to acknowledge the eight-member international expert panel drawn from various disciplines and organisations of whom Naomi Lee, Kassandra Karpathakis and Jon Herries agreed to be named.

Contributors SR drafted the initial version of the manuscript, which was then critically reviewed by the rest of the authors. All the coauthors have cited and approved the final version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests SR holds Directorship in Medi-AI. PM is cofounder of BrainX and BrainX Community. EC sits on the Board of Evidentli.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

ORCID iDs
Sandeep Reddy http://orcid.org/0000-0002-5824-4900
Enrico Coiera http://orcid.org/0000-0002-6444-6584
Piyush Mathur http://orcid.org/0000-0003-3777-8767

REFERENCES

1 Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med* 2019;112:22–8.
2 Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. *JAMA Netw Open* 2018;1:e182658.
3 Coiera E. The last mile: where artificial intelligence meets reality. *J Med Internet Res* 2019;21:e16323.
4 Reddy S, Allan S, Coghlan S, *et al*. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020;27:491–7.
5 Najafabadi MM, Villanustre F, Khoshgoftaar TM, *et al*. Deep learning applications and challenges in big data analytics. *J Big Data* 2015;2:1.
6 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health* 2020;2:e537–48.
7 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020;26:1351–63.
8 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* 2014;350:g7594.
9 Talmon JL, Ammenwerth E. "The declaration of Innsbruck": some reflections. *Stud Health Technol Inform* 2004;110:68–74.
10 Ammenwerth E, Gräber S, Herrmann G, *et al*. Evaluation of health information systems-problems and challenges. *Int J Med Inform* 2003;71:125–35.
11 Hörig H, Marincola E, Marincola FM. Obstacles and opportunities in translational research. *Nat Med* 2005;11:705–8.
12 Vis C, Bührmann L, Riper H, *et al*. Health technology assessment frameworks for eHealth: a systematic review. *Int J Technol Assess Health Care* 2020;36:204–16.
13 Yusuf M, Atal I, Li J, *et al*. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open* 2020;10:e034568.
14 Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform* 2021;28:e100251.
15 Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int J Med Inform* 2021;153:104510.
16 Schmidt J, Marques MRG, Botti S, *et al*. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater* 2019;5:83.
17 Terrin N, Schmid CH, Griffith JL, *et al*. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol* 2003;56:721–9.
18 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.
19 Reddy S. Artificial intelligence and healthcare—why they need each other? *Journal of Hospital Management and Health Policy* 2020:1–3.
20 Omoumi P, Ducarouge A, Tournier A, *et al*. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol* 2021;31:3786–96.
21 Kanagasingam Y, Xiao D, Vignarajan J, *et al*. Evaluation of artificial Intelligence-Based grading of diabetic retinopathy in primary care. *JAMA Netw Open* 2018;1:e182665.

SUPPLEMENT

Evaluation Framework Checklist and Scoring Sheet

Component	Initial Score	Weight	Component Score
1. Capability			
1.1. Objective **This subcomponent assesses whether the system has a clear objective i.e., stated contribution to a specific healthcare field. This subcomponent is scored on a scale of how clearly the objective is articulated.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	
The objective of the system is articulated	**1**		
The objective of the system and why the study is being conducted are clearly articulated	**2**		
The objective of the system and why the study is being conducted and how it adds to the body of knowledge in the domain are clearly articulated	**3**		

1.2. Dataset Source and Integrity

An AI system is only as good as the data it was derived from. If the training data does not reflect the intended purpose, the model predictions are likely to be useless or even harmful. This subcomponent evaluates the source of the data and the integrity of datasets used for training and testing the AI system including an appraisal of the representation of the target population in the data, coverage, accuracy and consistency of data collection processes and transparency of datasets. This subcomponent is scored on a scale of how well the dataset is described, how well the datasets fit with the ultimate objective and use case, and how credible/reliable the data source is. The subcomponent also considers when new data is acquired to train an embedded model that appropriate checks are undertaken to ensure integrity and alignment of data to previously used data

Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered, or acknowledged	**0**	**10**	
Non-representative training set, low quality or poorly defined collection protocol, dataset biased due to conflict of interests	**1**		
Training set is a sample from intended target population, appropriate data collection protocol, conflicts of interests not likely to drive conclusions	**2**		

Diverse training set of excellent coverage of all affected population segments, high quality and comprehensive data collection protocol, highly reputable and diverse data creators with no hidden agenda. When new data is acquired to train integrated models, appropriate check of data is undertaken as per above parameters	**3**		
1.3. Internal Validity **An internally valid model will be able to predict health outcomes reliably and accurately within a pre-defined set of data resources that were used wholly or partially when training the model. This includes the classical concept of goodness-of-fit, but also cross-validation schemes that derive training and tests sets from the same sources of data. Scoring is based on the size of the training data set with respect to the health care challenge, the diversity of the data to ensure good modelling coverage, and whether the statistical performance of the model (e.g., classification) is high enough to satisfy the requirements of clinical usefulness.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	
Small internal datasets, low statistical power, poorly defined or low-quality prediction target, inappropriate study design	**1**		

Adequate statistical power, sufficient data quality and accuracy, appropriate study design	**2**		
Extensive and representative internal datasets, high quality of measurements, careful consideration for confounders, gold standard prediction target	**3**		
1.4. External Validity **To qualify as external validation, we require that the external data used to assess AI system performance must come from substantially distinct external source that did NOT contribute any data towards model training. Examples of external data sources include independent hospitals, institutions or research groups that were not part of the model construction team or a substantial temporal difference between the training and validation data collections. The scoring is based on the size and diversity of the external data (if any) and how well the external data characteristics fit with the intended care recipients under the study objective.**			
Not applicable in this context	**NA**	**10**	
Not reported, considered or acknowledged	**0**		
Mismatched internal and external datasets (samples of different populations, data collected differently, outcomes defined differently), small or low-quality validation set	**1**		

Compatible internal and external datasets, sufficient statistical power, validation data from a different independent source from any training samples	**2**		
Extensive and multiple external validation datasets from diverse sources, excellent coverage of intended target population and real-world practice	**3**		
1.5. Performance Metrics **Performance metrics refers to mathematical formulas that are used for assessing how well an AI model predicts clinical or other health outcomes from the data. If the metrics are chosen poorly, it is not possible to assess the accuracy of the models reliably. Furthermore, specific metrics have biases, which means the use of multiple metrics is recommended for robust conclusions. This subcomponent examines whether performance measures relevant to the model and the results stated in the study are presented. These performance metrics can be classification or regression or qualitative metrics. This subcomponent is scored on a scale of how well the performance metrics fit the study and how reliable they are likely to be considering the nature of the health care challenge.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	
Only one formula used, uncertain accuracy of the performance measures, no replicates for cross-validation or similar framework	**1**		

Multiple metrics applied including those relevant for clinical practice, replicates or other techniques applied to consider reliability of performance measures	**2**		
Extensive analyses and benchmarking of relevant performance metrics across multiple datasets, careful experiments to verify the accuracy of performance values, mitigation of potential confounding factors. Further, when the model is updated, appropriate performance metrics are utilised to assess new outputs.	**3**		
1.6. Use Case **This subcomponent is seeking justification for the use of AI for the health need as opposed to other statistical or analytical methods. This tests if the application has considered the relevance and fit of the AI to the particular healthcare domain it is being applied to. This subcomponent is scored on a scale of how well the use case is stated.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**5**	
The case for AI use is vaguely formulated or not justified or inappropriate for the dataset, study design or objective	**1**		

The case for AI use is reasonable (against other alternatives) given the study design and results, and relevant for practice	**2**		
The case for AI use is clearly formulated or justified and appropriate for the dataset, study design or objective; AI is the best foreseeable solution to the health care challenge	**3**		
2. Utility			
2.1. Generalizability and Contextualization **The context of an AI application is defined here as the match between the model performance, expected features, characteristics of the training data and the overall objective. In particular, biases or exacerbation of disparities due to underrepresentation or inappropriate representation due to the availability of datasets used both in training and validation can have an adverse effect on the real-world utility of an AI model. This subcomponent is scored based on how well it is expected to perform on the specific groups of people it is most intended for.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	

Limited diversity of people or insufficient clinical description to confirm if representative of intended care recipients	**1**		
Representative of intended care recipients, appropriate eligibility criteria and baseline characteristics for practical applications	**2**		
Real-world setting of diverse care patients, excellent coverage of multiple ethnic, socioeconomic and identity groups	**3**		
2.2. Safety and Quality **It is critical that AI models being deployed in healthcare, especially in clinical environments, are assessed for their safety and quality. Appropriate consideration should be paid to the presence of ongoing monitoring mechanisms in the study, such as adequate clinical governance that will provide a systematic approach to maintaining and improving the safety and quality of care within a healthcare setting. This subcomponent is scored based on the strength of the safety and quality process and how likely it is to ensure safety and quality when AI is applied in the real-world.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered, or acknowledged	**0**	**10**	
Safety measures and quality controls presented are likely to be inadequate	**1**		

Reasonable steps taken to adopt safety and quality processes but an ongoing sustainable governance framework is not in place	**2**		
Careful consideration and testing of possible adverse impacts, continuous safety and quality monitoring mechanisms are in place and active	**3**		
2.3. Transparency **This subcomponent assesses the extent to which model functionality and architecture is described in the study and the extent to which decisions reached by the algorithm are understandable (i.e., black box or interpretable). Important elements are the overall model structure, the individual model components, the learning algorithm, and how the specific solution is reached by the algorithm. This subcomponent is scored on a scale of how transparent, interpretable and reproducible the AI models are, given the information available.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	
Basic algorithm concept published in a peer-reviewed journal or a technical report, sufficient technical description to reproduce essential components of the pipeline	**1**		

Algorithm is published and source code or software tools are available for inspection, sufficient statistical analyses to describe the patterns in the data that are likely driving the prediction outcomes	**2**		
Fully open-source project with public access to training data; the process by which the algorithm derives predictions from data is analysed statistically, carefully documented and visualized/presented in ways that are easy to interpret by humans	**3**		
2.4. Privacy **This subcomponent covers personal privacy, data protection and security. This subcomponent is ethically relevant to the concept of autonomy/self-determination, the right to control access to and use of personal information, and the consent processes used to authorize data uses. This subcomponent is scored on the extent of consideration of privacy aspects including consent by study subjects, the strength of data security and data life cycle throughout the study itself and consideration for future protection if deployed in the real-world.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	

Informed consent when applicable, personal and sensitive data protected from unauthorized access	**1**		
Anonymised research data, study protocol conforms to international standard on privacy/ethics, approved by local ethics committee, informed consent when applicable	**2**		
Privacy by design, peer-reviewed study protocol approved by ethics committee and satisfies the highest international standards, facility to opt-out at any time, past and future data securely stored with database usage, disposal plan in place	**3**		
2.5. Non-Maleficence **This subcomponent refers to the identification of actual and potential harms caused by the AI and actions to avoid foreseeable or unintentional harms. Harms to individuals may be physical, psychological, emotional, economic. Harms may affect systems/organizations, infrastructure and social wellbeing. This subcomponent is scored on the extent to which potential harms of the AI are identified, quantified and the measures taken to avoid harms and reduce risk.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	

Potential harms are acknowledged (e.g., misdiagnoses, bodily harms, discrimination, loss of trust, loss of skills) but effective mitigation not presented	**1**		
Potential harms are discussed (see above) and specific actions to minimise harms are presented	**2**		
Systematic approach to minimise harms, demonstrated effectiveness of risk management, independent and appropriate audits and governance	**3**		
3. Adoption			
3.1. Use in a Healthcare Setting **As discussed earlier, many AI systems have been developed in controlled environments or in-silico, but there is a need to assess for evidence of use in real world environments and integration of new AI models with existing information systems. This subcomponent is scored according to the extent to which the model has been adopted by and integrated into 'real world' healthcare services i.e., healthcare settings beyond the test site.** This subcomponent also considers the applicability of the system to end-users, both clinicians and administrators, and the beneficiaries of the system, patients as part of the evaluation.			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	

Vague or insufficient description of practical integration into hospital or other care environments beyond the test site	**1**		
Detailed and convincing plan to integrate methodology into specific clinical/administrative workflows beyond the test site	**2**		
Practical demonstration of how method is integrated into workflows across institutes, governance oversight clearly defined and in place beyond the test site and assessment of the benefits or impact on end users and intended beneficiaries	**3**		
3.2. Technical Integration **This subcomponent evaluates how well the AI systems integrate with existing clinical/administrative workflows outside of the development setting, and their performance in such situations. In addition, the subcomponent includes reporting of integration even if the model performs poorly. This subcomponent is scored on a scale of how well the integration aspects of the model are anticipated and if specific steps to facilitate practical integration have been taken.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**10**	

Insufficient model performance for purpose, poor fit, presence of artefacts	**1**		
Solid evidence for sufficient model performance and accuracy for purpose across multiple metrics	**2**		
Extensive evidence for sufficient and stable model performance for purpose across multiple settings and datasets, highly likely or demonstrated performance in intended real-world environment	**3**		
3.3. Number of Services **Many AI in healthcare studies are based on single site use without evidence of wider testing or validation. In this subcomponent, we review reporting of wider use. This subcomponent is scored on a scale of how well the use of the model across multiple healthcare organizations is described.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**5**	
Deployed or piloted in one healthcare service or entity	**1**		

Established long-term deployment in a single large healthcare service or multiple national deployments across different services	**2**		
International deployment across multiple healthcare services and entities with a demonstrated record of success	**3**		
3.4. Alignment with Domain **This category considers how much of information about the alignment and relevance of the AI system to the healthcare domain and its likely long-term acceptance are reported. In other words, the model is assessing the benefits of the AI model to the particular medical domain the model is being applied to. This again relates to the translational aspects of the AI model. This subcomponent is scored on a scale of how well the benefits of the AI model to the medical domain are articulated.**			
Not applicable in this context	**NA**	**NA**	**NA**
Not reported, considered or acknowledged	**0**	**5**	
Poor justification for clinical benefit due to lack of societal need, clinical evidence or scientific rationale	**1**		
Sound justification for added clinical benefit from AI, good methodological fit between chosen approach and problem to be solved	**2**		

Objective is a direct response to pressing medical need, high likelihood of transformative impact, solution difficult without AI	**3**		