BMJ Health & Care Informatics

Clinician checklist for assessing suitability of machine learning applications in healthcare

Ian Scott,[1,2] Stacy Carter,[3] Enrico Coiera[4]

To cite: Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform* 2021;**28**:e100251. doi:10.1136/bmjhci-2020-100251

Received 03 October 2020
Accepted 12 January 2021

[1]Internal Medicine and Clinical Epidemiology, Princess Alexandra Hospital, Brisbane, Queensland, Australia
[2]School of Clinical Medicine, Univeristy of Queensland, Brisbane, Queensland, Australia
[3]Australian Centre for Health Engagement Evidence and Values, University of Wollongong, Wollongong, New South Wales, Australia
[4]Centre for Health Informatics, Macquarie University, Sydney, New South Wales, Australia

Correspondence to
Professor Ian Scott;
ian.scott@health.qld.gov.au

ABSTRACT

Machine learning algorithms are being used to screen and diagnose disease, prognosticate and predict therapeutic responses. Hundreds of new algorithms are being developed, but whether they improve clinical decision making and patient outcomes remains uncertain. If clinicians are to use algorithms, they need to be reassured that key issues relating to their validity, utility, feasibility, safety and ethical use have been addressed. We propose a checklist of 10 questions that clinicians can ask of those advocating for the use of a particular algorithm, but which do not expect clinicians, as non-experts, to demonstrate mastery over what can be highly complex statistical and computational concepts. The questions are: (1) What is the purpose and context of the algorithm? (2) How good were the data used to train the algorithm? (3) Were there sufficient data to train the algorithm? (4) How well does the algorithm perform? (5) Is the algorithm transferable to new clinical settings? (6) Are the outputs of the algorithm clinically intelligible? (7) How will this algorithm fit into and complement current workflows? (8) Has use of the algorithm been shown to improve patient care and outcomes? (9) Could the algorithm cause patient harm? and (10) Does use of the algorithm raise ethical, legal or social concerns? We provide examples where an algorithm may raise concerns and apply the checklist to a recent review of diagnostic imaging applications. This checklist aims to assist clinicians in assessing algorithm readiness for routine care and identify situations where further refinement and evaluation is required prior to large-scale use.

As a subset of artificial intelligence, machine learning (ML) is being used to create algorithms to screen and diagnose disease, prognosticate, and predict response to clinical interventions (box 1). Deep learning (DL), which uses massive artificial neural networks, has been responsible for much recent progress in ML. More than 150 clinical DL algorithms have now passed proof-of-concept phase,[1] and over 50 have been approved for routine use by the US Food and Drug Administration.[2]

However, before adopting algorithms into routine care, practising clinicians will seek reassurance from their professional bodies and healthcare institutions about their validity, utility, feasibility, safety and ethical use. Amidst the hype and opaque nature of many ML applications, and contestable claims of superior performance of some algorithms compared with clinical experts,[3] clinicians need to have some understanding of how algorithms are developed and how to assess their clinical worth.

Recent commentaries have identified several important challenges relating to ML applications in healthcare which end-users need to be aware of when deciding whether to adopt them into routine care.[4–8] We developed a checklist that reflect these challenges in a manner suitable to the needs and training of practising clinicians. It contains questions clinicians should ask of algorithm developers, vendors and implementers. In so doing, we recognise that, as non-experts in ML, clinicians cannot be expected to demonstrate mastery over what can be highly complex statistical and computational concepts. In seeking answers to certain questions, they may need to depend on the expertise of data scientists or health informaticians. In formulating the checklist, we made reference to recent narrative reviews,[1 9–12] a report from the US National Academy of Medicine,[13] and recent studies (from 2000) published in PubMed using search terms 'ML,' 'DL' and related synonyms.

Q1. WHAT IS THE PURPOSE AND CONTEXT OF THE ALGORITHM?

Algorithm development should be driven by a clinical need or 'pain point', not what is simply technically feasible by virtue of available data. Clinicians should ask if, at the design phase, developers collaborated with end-users in agreeing: (1) the specific clinical task or function of the algorithm (diagnosis, prognostication, treatment response); (2) the target population and clinical setting and (3) the intended method of algorithm implementation.[4]

Q2. HOW GOOD WERE THE DATA USED TO TRAIN THE ALGORITHM?

Algorithms can only be as good as the data they were trained on, and that data need to be

easily accessible where the algorithm is to be used, easily migrated into different computer programmes (interoperable), and able to be stored and reused.

Q2a. To what extent were the data accurate and free of bias?

In assuring algorithm accuracy, clinicians should confirm that datasets used to train an algorithm were of high quality, representative of the population of interest, derived from reliable sources and had minimal missing data.[14] Many algorithms use transactional data from electronic medical records (EMRs) or administrative datasets—typically of poorer quality than clinical registry and trial datasets. However, given their extensive coverage of clinical care and their availability, such data will continue to be used. However, clinicians should note that incomplete, inaccurate, poorly described or incorrectly labelled data are more likely to introduce error.

Even more important are systematic biases in what data were collected, how and on whom. Some variables highly relevant to clinical outcomes (ancestry, language, socioeconomic status, laboratory tests, health-related circumstances, such as substance abuse, physical activity and homelessness) may not be routinely captured.[6] For example, a cardiovascular risk prediction algorithm was inaccurate in marginalised populations because training data were never obtained from them (selection bias).[15] An algorithm predicting survival of post-menopausal women using electrocardiographic markers, clinical characteristics and demographic variables performed worse than conventional Framingham scores, partly because it lacked important blood test results (measurement bias).[16] Recent research detected racial bias in an algorithm that could potentially affect millions of patients.[17]

Clinicians need to ask: what were the criteria for selecting patients for the training dataset, how many were screened and included, were all relevant baseline characteristics measured in all individuals, and what was done

Box 1 Machine learning (ML)—background concepts and examples

ML is the process whereby advanced computer programs (machines), often with minimal human instruction, process often huge datasets (big data), potentially from many sources, to discern patterns and associations which are then used to iteratively encode (or learn) a process or system model (algorithm). This algorithm, when applied to new data, aims to produce a prediction or outcome more quickly and accurately than clinical experts, devoid of errors due to human cognitive bias and fatigue.

Algorithms are developed (or trained) using training datasets derived from medical imaging devices, electronic medical records, administrative datasets or wearable biosensors. The trained algorithms may be tuned and then tested on samples of the training datasets to gauge accuracy and reproducibility, and then validated on new unseen datasets in assessing their generalisability to new populations and settings.

Types of ML

- Supervised learning maps input data from a training set of labelled (or known) examples to generate a model which can be applied to new data in making predictions. As the examples are already known, the model learns 'under supervision'. Supervised learning is used for classification (eg, discriminating between different items, categories or subgroups in making a diagnosis) and regression (prediction) (eg, estimating the likelihood of a future clinical event).
- Unsupervised learning uses input data from unlabelled examples and groups them according to some attribute (or pattern) of shared commonality. Unsupervised learning is used for: clustering, that is, identifying and characterising clusters of variables that appear to share latent similarities; and anomaly detection, that is, identifying unusual patterns of outlier or dissimilar values for different variables. An example is where clinical and genetic data from thousands of patients with a certain diagnosis, and who have been managed in different ways, are processed in identifying genotypic or phenotypic features associated with favourable or unfavourable response to certain treatments.
- Reinforcement learning processes dynamic data that is constantly changing and where the algorithm adapts to change and learns an optimised set of rules for achieving a goal or maximising an expected return (or reward) by a process of trial and error. Model behaviour is 'reinforced' by the level of reward achieved. Examples may include controlling an artificial pancreas system to fine-tune the measurement and delivery of insulin to patients with diabetes, or adjusting ventilator and vasopressor infusion rates in seriously ill patients in intensive care units.

Classes of ML algorithms

There are more than 20 different classes of ML algorithms; the following are the most commonly encountered.

- Artificial neural networks are non-linear algorithms loosely inspired by human brain synapses, with the most common being convoluted neural networks (or deep learning). These networks comprise input nodes, output nodes and intervening or hidden layers of nodes, which may number up to 100. Each node within a layer involves two or more inputs and applies an activation and weighting function to produce an output which serves as the input data for the next layer of nodes. In deep learning, data from imaging devices is passed through successive layers of nodes which convolute (transform) and pool the data and extract high order features such as contrast, colour, shapes, edges and patterns. These feature maps are successively pooled to produce the final outputs.
- Support vector machines (SVMs) transform input data into two classes or categories by choosing the boundary or widest plane (or support vector) that separates them to the maximal degree. SVMs can map examples to other dimensions which have non-linear relationships, and by transforming low dimensional input data into high-dimensional space using mathematical tools (kernel functions), they can separate such examples linearly by determining a hyperplane as the decision surface.
- Decision trees choose a series of sequential branching decisions on features in the training data which map the features to a known outcome with the most accuracy. They may use naïve Bayesian methods which assign pretest probabilities or prevalence to certain features and assume all features are independent of one another, or use random forests which adopt a completely random order of branching steps in a subset of training examples. Similar to SVMs, the goal is to optimally separate the classes in training examples.

to account for missing data or time varying confounders, such as downstream clinical management decisions? Because algorithms can learn, automate and accentuate existing biases in training datasets, thereby worsening healthcare inequities,[18] strategies for mitigating these biases during the training process[19] should be stated.

Q2b. Were data labelled correctly?

Supervised learning, currently the most common type of ML, may require training data to be labelled with the category or class of interest. For example, a retinal image might be labelled as showing diabetic retinopathy, where diabetes can be confirmed by a glycosylated haemoglobin test, but diagnosing retinopathy relies on subjective judgement of ophthalmologists. In avoiding algorithms developed using unreliable labels, clinicians should ask what reference standards (or 'ground truths') were used in deciding whether, in this case, diabetic retinopathy was the correct diagnosis. The ideal standard is often consensus adjudication by panels of expert clinicians, blind to algorithm predictions and given sufficient time and clinical information—reflecting normal clinical practice—to make well-considered predictions of whether a particular abnormality is present, absent or indeterminate.[20]

Q2c. Were the data standardised and interoperable?

Most algorithms are initially programmed to have data presented to them in a format (or 'common data model') that accords with a specific data standard. Imaging data are typically well standardised and interoperable using the Digital Imaging and Communications in Medicine and Picture Archiving and Communication System standards. However, for structured data within clinical

Box 2 Performance measures for machine learning algorithms

Area under receiver operating characteristic curve (AUROC)

For binary outcomes involving numerical samples (such as disease or event present or absent), the receiver operating characteristic (ROC) curve plots the true positive (TP) rate (sensitivity) against the false positive rate (1 minus specificity). An AUROC of 1.0 represents perfect prediction; an AUROC equal to or above 0.8 is preferred.

For binary outcomes involving imaging data, a modification of the ROC is the free-response ROC, or FROC* where a FROC curve comprising a 45º diagonal line indicates the algorithm is useless, while the steeper and more convex the slope of the curve, the greater the accuracy.

In situations where outcomes are not binary and multidimensional, or where data are highly skewed with disproportionately large numbers of true negatives, other methods such as the volume under the surface of the ROC curve and false discovery rate-controlled area under the ROC curve have been suggested; values equal to or above 0.8 are again preferred.**

Confusion matrix

A confusion matrix is a contingency table which yields several metrics, with optimal performance represented by values approaching 100% or 1.0.

- Positive predictive value (PPV) or precision: the proportion of positive cases that are TP rather than false positives (FP): PPV=TP/TP +FP.
- Negative predictive value (NPV): the proportion of negative cases that are true negatives (TN) rather than false negatives (FN): NPV=TN/TN +FN.
- Sensitivity (Sn) or recall: the proportion of TP cases that are correctly identified: Sn=TP/TP+FN.
- Specificity (Sp): the proportion of true negative (TN) cases which are correctly identified: Sp=TN/TN+FP.
- Accuracy: the proportion of the total number of predictions that are correct: TP+TN/TP+FP+TN+FN.
- F1 score: this measure represents the harmonic mean of precision (or PPV) and recall (sensitivity) in which both are maximised to the largest extent possible, given that one comes at the expense of the other. It is reported as a single score from 0 to 1 using the formula: 2 x TP/(2 x TP+FP+FN). The higher the score, the better the performance.
- Matthew's correlation coefficient: This coefficient takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes: TP x TN – FP x FN/√ (TP +FP) (TP+FN) (TN+FP) (TN+FN). A coefficient of +1 represents a perfect prediction, 0 no better than random, and –1 total disagreement between prediction and actual outcome.

Precision-recall (PR) curve

The PR curve is a graphical plot of PPV (or precision) against sensitivity (or recall) to show the trade-off between the two measures for different feature (or parameter) settings. The area under the PR curve is a better measure of accuracy for classification tasks involving highly imbalanced datasets (ie, very few positive cases and large numbers of negative cases). An area under the PR curve (AUPRC) of 0.5 is preferred. Ideally, algorithm developers should report both AUROC and AUPRC, along with figures of the actual curves.

Regression metrics

Various metrics can be used to measure performance of algorithms performing regression functions (ie, predicting a continuous outcome). They include mean absolute error (mean of the absolute differences between actual and predicted values), mean squared error (calculated by summing the differences between actual and predicted values, squaring the results, and dividing by the total number of instances) and root mean squared error (standard deviation of all errors). In all cases, values closer to 0 indicate better performance.

Another commonly used metric is the coefficient of determination (R^2), which represents how much of the variation in the output variable (or Y—dependent variable) of the algorithm is explained by variation in its input variables (X—independent variables). An R^2 of 0 means prediction is impossible based on input variables and R^2 of 1 means completely accurate prediction with no variability. Generally R^2 should be above 0.6 for the algorithm to be useful.

*See Moskowitz CS. Using free-response receiver operating characteristic curves to assess the accuracy of machine diagnosis of cancer. *JAMA* 2017;318:2250–2251.

**See Yu T. ROCS: Receiver operating characteristic surface for class-skewed high-throughput data. *PLoS One* 2012;7:e40598.

records, different standards exist, for example, Systematised Nomenclature of Medicine-Clinical Terms[21] or the Observational Medical Outcomes Partnership standard.[22] In mapping data from one standard to another, the more mapping required, the greater the cost and risk of inducing errors.[23] Fortunately, the HL7-Fast Healthcare Interoperability Resources is emerging as a robust, standard-agnostic messaging system which facilitates data migration with minimal need for mapping.[24] Mapping unstructured, free-text clinical data is more challenging, although natural language processing algorithms can map words to clinical concepts.[25] Clinicians should ask if significant mapping work is required to meet local data standards before implementing an algorithm, and inquire into the costs and risks of doing so.

Q3. WERE THERE SUFFICIENT DATA TO TRAIN THE ALGORITHM?

In general, the more complex the algorithm, in having to make more distinctions between a larger number of different things, the more data required. Convolutional neural networks used to process medical images or text or huge numerical datasets may require many thousands of training examples.[26] However, methods for determining a priori just how many examples are required are yet to be agreed.[27] If more data continues to improve algorithm performance, more data should be supplied. Clinicians should be informed of how much data were used, how that sample size decision was reached, and what techniques (such as feature engineering and regularisation procedures) were used to deal with data of high dimensionality (ie, possessing many different attributes, as in imaging data) or of limited availability, as these all bear on algorithm performance.[28]

Box 3 Ethical, legal and social issues of using algorithms[61–66]

- How were consent issues handled in collecting data used for algorithm training and validation?
- Who owns, or has stewardship of, the data and determines how it is to be used in training and testing of algorithms?
- How are data confidentiality and patient privacy ensured when data is stored (in the cloud) and used and shared across different platforms?
- How much responsibility for care should clinicians be expected to assume when using algorithms they cannot control or explain?
- Who carries liability if patients are injured by a faulty or misapplied algorithm (developers who trained and tested the algorithm, vendors who integrated the algorithm into electronic medical records or imaging software, or clinicians using the algorithm to make decisions)?
- Who takes responsibility for postimplementation monitoring of the safety and efficacy of an algorithm throughout its life cycle, and determine when an algorithm needs updating, retraining or even withdrawal because of emerging inaccuracies?
- Will the majority of clinicians (and patients) be literate enough to understand how, when and in whom machine learning algorithms are safe and effective to use?
- How equitable and inclusive are the algorithms? Is there risk of a digital divide between healthcare institutions (and their catchment populations) who can or cannot deploy or access algorithm systems (for various reasons)?
- Who might have conflicts of interest in developing, disseminating, using or advocating a particular algorithm?
- Who owns the intellectual property pertaining to an algorithm; who owns the patent rights; who and what factors determine whether an algorithm is able to be commercialised for profit?

Q4. HOW WELL DOES THE ALGORITHM PERFORM?

Just as with a diagnostic test or a prediction rule, clinicians should be told the accuracy and reproducibility of algorithm outputs. A process of internal (or in-sample) validation should have tested and refined the algorithm on datasets resampled from the original training datasets,[29] either by bootstrapping (multiple sampling in random order) or cross-validation (datasets segmented into different testing sets multiple times [or 'folds'], hence the term k-fold cross-validation where k=number of folds, usually 5 or 10).

This is followed by a process of external (out-of-sample) validation on previously unseen data, preferably taken from a temporally or geographically different population. This step, which is often omitted, is crucial as it often reveals overfitting, where the algorithm has learnt features of the training dataset too perfectly, including minor random fluctuations, and consequently, may not perform well on new datasets. For classification tasks which are most common, metrics of discrimination should be reported (box 2), and chosen sensitivity/specificity thresholds justified in maximising clinical utility.[30] For regression-based prediction tasks, clinicians should ask if an algorithm performs better than existing regression models, in case it may not,[31] and ask if replication studies of the same algorithm by independent investigators have yielded the same performance results.[32]

Q5. IS THE ALGORITHM TRANSFERABLE TO NEW CLINICAL SETTINGS?

A crucial question for clinicians is whether the algorithm performs equally well across a range of new clinical settings and, if not, can the algorithm be retuned or recalibrated using local data to account for differences in population characteristics, type or reporting formats of imaging devices, or care protocols.[33 34] For example, a DL system for interpreting thyroid ultrasound images in detecting cancers saw sensitivity drop from 92% (human equivalent) to 84% (below human), with no change in specificity, when applied to different hospitals.[35] An algorithm used to diagnose pneumonia on chest X-rays in one hospital system failed to generalise to radiographs from another hospital system, due to differences in prevalence of pneumonia between populations[36] (class imbalance). Differences in illness severity can also degrade performance of algorithms trained on more severely diseased populations when applied to those with mild or moderate

Table 1 Application of the checklist

Liu *et al*[67] analysed 82 studies published between January 2012 and June 2019 which compared diagnostic performance of deep learning algorithms and healthcare professionals based on medical imaging for 17 different clinical conditions. The authors extracted diagnostic accuracy data and constructed contingency tables to derive the measures of interest. In generating responses to each item on the checklist, we used information stated in the review or, if certain information was missing, retrieved from the individual full-text articles.

Item	Response
1. What is the purpose of the algorithm?	Objective and context of the algorithms were adequately stated in included studies.
2a. How good were the data used to train the algorithm? 2b. To what extent were the data accurate and free of bias? 2c. Were the data standardised and interoperable?	26 studies (32%) did not report patient inclusion criteria; 33 studies (40%) did not report exclusion criteria; 30 studies (37%) did not report age and 43 studies (52%) did not report sex. 72 studies (88%) used retrospectively collected data from historical routine care (48 studies) or open source (24 studies) registries which are rarely quality controlled for images or accompanying labels, and in which population characteristics are either not collected or inaccessible; only 10 studies (12%) used prospectively collected data specific to a research setting. 26 studies (32%) excluded low-quality images; 18 (22%) retained low-quality images; 38 (46%) did not report this. The extent of missing data, and how this was handled, was poorly reported in all studies. All data used in 36 studies (44%) were obtained at a single hospital or medical centre. The extent to which data were standardised and rendered interoperable across sites in multisite studies was not reported in any study.
3. Were there sufficient data to train the algorithm?	57 studies (69%) did not report the number of participants represented by the training data; in remaining studies, the numbers ranged from 40 to 200 000. No study pre-specified a sample size.
4. How well does the algorithm perform?	For internal validation, 22 studies (27%) used resampling methods, 29 studies (35%) used random split sampling, 1 study (1%) used stratified random sampling, and 30 studies (37%) did not report any form of internal validation. 69 studies (84%) provided adequate data to construct contingency tables. In these studies sensitivity ranged from 9.7% to 100.0% (mean±SD 79.1%±0.2%); specificity ranged from 38.9% to 100.0% (mean±SD 88.3%±0.1%). Only 12 studies (14.6%) reported cut-points for determining sensitivity and specificity for which no justification was provided. The same reference standard was used across internal validation datasets in 61 studies (74%). Reference standards varied widely according to target condition and imaging modality. More rigorous expert group consensus standards were used in 66 studies (80%); remaining studies relied on single expert consensus (n=1), existing clinical care notes or imaging reports or existing labels (n=11), clinical follow-up (n=9), surgical confirmation (n=2), another imaging modality (n=1) and laboratory testing (n=3). No comments were made about outlier studies although AUROC curves depicted within the review clearly indicated there were such studies. Only 25 of 82 studies (36%) performed external validation. In these studies, the pooled sensitivity was 88.6% (95% CI 85.7 to 90.9) and pooled specificity was 93.9% (95% CI 92.2 to 95.3). Studies were inconsistent in their use of the term 'validation' as it applied to testing datasets; there was often lack of transparency as to whether testing sets were truly independent of training sets.
5. Is the algorithm transferable to new clinical settings?	Only 9 studies (11%) assessed algorithm performance in real-world contexts where clinicians received additional clinical information alongside the image, rather than just view the image in isolation.
6. Are the outputs of the algorithm clinically intelligible?	81 studies (99%) used artificial or convoluted neural networks; 1 study did not report algorithm architecture. Only 32 studies (39%) provided a heat map of salient features.
7. How will this algorithm fit into and complement current workflows?	No studies reported how their algorithms impacted real-world clinical workflows. In one study which compared algorithm performance among pathologists simulating normal workflows (ie, imposed time constraints) with that of a single pathologist with no time constraint, the AUROC were the same (0.96).*
8. Has use of the algorithm been shown to improve patient care and outcomes?	None of the algorithms in these studies have been subjected to clinical trials aimed at demonstrating improved care or patient outcomes.
9. Could the algorithm cause patient harm?	No comments were made about potential harms.
10. Does use of the algorithm raise ethical, legal or social concerns?	No comments were made about any such concerns.

*Bhteshami Bejnordi BE, Veta M, van Diest PJ, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318(22):2199–2210.
AUROC, area under receiving operator characteristic curve.

disease (spectrum bias). Variations in data quality, clinical actions included in the algorithm (causality leakage) or classification of outcomes (label leakage) can also affect local performance. While methods are emerging to minimise these problems,[37 38] clinicians should ask if the algorithm is applicable to their local setting, and whether it may need recalibration using local data.

Q6. ARE THE OUTPUTS OF THE ALGORITHM CLINICALLY INTELLIGIBLE?

Clinicians may not trust 'black box' algorithms which produce diagnoses or predictions in difficult-to-interpret formats, or provide little explanation of how these outputs were generated, especially those that appear counterintuitive. For the former, output formats may need to be customised to those that facilitate rapid clinical interpretation.[39] For the latter, decision trees and Bayesian networks are readily explainable in how they model causality, but data-driven methods, such as DL do so only implicitly, and may confuse association with causation, leading in some cases to clinically incorrect inferences. For example, an algorithm predicting low-risk patients with pneumonia who could be safely discharged from hospital was found to have incorrectly classified high risk asthmatic patients as low risk,[40] unaware that, by being routinely admitted to intensive care units, such patients had better survival. Another algorithm for detecting pneumothoraces on chest X-rays was trained on films taken after chest tube insertion, thus learning to identify chest tubes rather than pneumothoraces.[41]

In affording clinicians a better understanding of how algorithms generate their conclusions, various software tools can identify the features an algorithm chose as being critical in forming its predictions (eg, Local Interpretable Algorithm-Agnostic Explanations and Shapley Values in Machine Learning (SHAP)). These programmes can produce saliency or heat maps, pinpointing the exact areas and features in an image the algorithm has decided are abnormal,[42] and deconvolution graphs, highlighting the variables the algorithm regards as being most informative in predicting risk.[43]

Q7. HOW WILL THIS ALGORITHM FIT INTO AND COMPLEMENT CURRENT WORKFLOWS?

The utility of any algorithm in routine practice depends greatly on its 'fit' into clinical work and its impact on clinician time, efficiency and cognitive load. For example, in detecting metastatic breast tumours in sentinel lymph node biopsies, highlighting only the most suspicious regions expedited image review by pathologists, while showing raw algorithm predictions of each region of the image slowed them down.[44] Research into the ergonomics of using algorithms in routine clinical care is currently very limited, especially as the effort required for successful implementation can vary widely across even similar healthcare organisations because of subtle variations in workflows, tasks and patient needs.

Automating entry of imaging or EMR data into algorithms which self-activate in response to specific orders or requests can potentially help generate timely, actionable outputs.[45 46] The absence of such automation may simply increase burden of work on users, causing them to devise workarounds to avoid using an algorithm or abandoning it altogether.[47] Clinicians should therefore consider: (1) the exact point in the clinical trajectory where the algorithm will be applied; (2) the way the algorithm would actually be implemented in a specific clinical setting, and the technical and staff training effort required; (3) the resulting workflow changes and (4) the level of use the algorithm would likely receive from its intended users.

Q8. HAS USE OF THE ALGORITHM BEEN SHOWN TO IMPROVE PATIENT CARE AND OUTCOMES?

An algorithm will likely be ignored if clinicians do not perceive it as improving patient care and outcomes, either because the current human system is already optimal, or the algorithm is too far removed from critical decision points. Screening applications in otherwise healthy populations,[48] in whom inaccurate algorithms may cause significant harm, warrant careful attention. Rigorous clinical impact studies of DL algorithms are, to date, infrequent,[3 49] most are uncontrolled pre-post or cohort studies, and clinical effects are sometimes very marginal.[50] Ideally, the algorithm should be implemented and tested for utility in pilot studies in 'silent' mode (real-time predictions exposed to clinical experts but not acted on, so errors can be identified), then tested for efficacy in prospective clinical trials, and finally assessed for effectiveness and cost-effectiveness in large-scale studies.[51 52] Importantly, more rigorous testing should apply as algorithms move from narrow diagnostic imaging applications to more complex therapeutic scenarios, and from assistive applications informing decisions to fully automated applications determining patient management independently of clinicians.

Q9. COULD THE ALGORITHM CAUSE PATIENT HARM?

Poorly calibrated algorithms applied to insurance risk, employability and other forms of social profiling have generated false and detrimental predictions.[53] ML algorithms have generated unsafe drug recommendations in oncology.[54] Algorithms can quickly become inaccurate or out of date, and need retraining due to changes in background characteristics, exposures or outcomes of patient populations (distributional shifts), unanticipated changes in clinical practices or patient behaviour (calibration drift), and persistence of outmoded clinical technologies.[55 56] Even changes in clinical care due to algorithm implementation can, in itself, cause data shifts.[57] Adversarial cyber attacks can corrupt either the datasets or the computer programmes underpinning

the algorithm, with effects potentially indiscernible to humans.[58] Automation bias may see clinicians become deskilled over time by over-reliance on algorithms,[59] leading to misdiagnoses and inappropriate therapeutics. Algorithms may encourage overdiagnosis by detecting subclinical anomalies that prompt unwarranted intervention.[60] Algorithms are unlikely to recognise when their outputs are false or affected by bias, and hence clinician must continue to question counter-intuitive or potentially harmful predictions.

Q10. DOES THE ALGORITHM RAISE ETHICAL, LEGAL OR SOCIAL CONCERNS?

Several contestable and intertwined ethical, legal and social issues are raised in using algorithms (box 3)[61–63] that clinicians need to consider, particularly personal liability for algorithm-induced harm[64] and blatant misuse of patient data that breaches privacy rules[65] enshrined in the US Health Insurance Portability and Insurance Act, the UK Data Protection Bill and the European General Data Protection Regulation. Numerous reports[66] provide guidance around clinician and patient autonomy, data privacy and governance processes, potential commercial conflicts of interest, openness (open data sets, methods and source code) and transparency, non-discrimination and fairness.

Application of the checklist

As a test of its potential utility, we applied our checklist to a recent systematic review of studies comparing accuracy of diagnostic imaging algorithms with that of clinical experts[67] (table 1). While this exercise did not target a single algorithm, which may be a limitation, our impression was that many studies demonstrated shortcomings for virtually every question—a problem which recently issued reporting guidelines for ML studies[68 69] will hopefully improve. In the meantime, our checklist may serve to protect clinicians from premature adoption of algorithms of uncertain worth.

CONCLUSION

Most clinicians will likely see ML algorithms increasingly used to augment their decision making. Image-intensive disciplines will likely see major reconfiguration of roles as algorithms are adopted to improve diagnostic accuracy. Algorithms will not replace clinicians, but clinicians who use well-designed and validated algorithms appropriately may replace those who do not. Clinicians need to be able to judge algorithm readiness for use and identify situations where further refinement and evaluation are needed prior to large-scale use.

Acknowledgements The authors wish to acknowledge Professor Jenny Doust (Bond University, Gold Coast, Australia), Nazanin hahreman-Falconer (Pharmacy PhD student, Princess Alexandra Hospital, Brisbane, Australia) and Ahmad Abdel-Hafez (Principal Data Analyst, Clinical Informatics, Metro South Health Service, Brisbane, Australia) for their comments on previous drafts.

Contributors IAS conceived the idea, undertook the primary research and drafted the first manuscript. EC and SC critically reviewed the manuscript, added new themes and references and assisted in redrafting the manuscript. IAS is the guarantor who accepts full responsibility for the finished article and had access to any data from literature searches.

Competing interests None declared.

Patient and public involvement statement We did not involve patients or the public in this work which comprised secondary research of published literature. There are no individual patient results to disseminate to these groups.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58.
2 US Food and Drug Administration. Fda cleared AI algorithms. data science Institute. Available: https://www.acrdsi.org/DSI-Services/FDA-cleared-ai-algorithms [Accessed 9 Sep 2020].
3 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.
4 Gilvary C, Madhukar N, Elkhader J, *et al*. The missing pieces of artificial intelligence in medicine. *Trends Pharmacol Sci* 2019;40:555–64.
5 Lindsell CJ, Stead WW, Johnson KB. Action-Informed artificial Intelligence-Matching the algorithm to the problem. *JAMA* 2020;323:2141.
6 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
7 Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: Humanism and artificial intelligence. *JAMA* 2018;319:19–20.
8 Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018;320:2199–200.
9 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
10 Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med* 2019;25:24–9.
11 He J, Baxter SL, Xu J, *et al*. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30–6.
12 Liu Y, Chen P-HC, Krause J, *et al*. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806–16.
13 Matheny MS, Israni T, Ahmed M, *et al*, eds. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. NAM Special Publication*. Washington, DC: National Academy of Medicine, 2019.
14 Goldstein BA, Navar AM, Pencina MJ, *et al*. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198–208.
15 Gijsberts CM, Groenewegen KA, Hoefer IE, *et al*. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One* 2015;10:e0132321.
16 Gorodeski EZ, Ishwaran H, Kogalur UB, *et al*. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the women's health Initiative. *Circ Cardiovasc Qual Outcomes* 2011;4:521–32.
17 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
18 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
19 Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019;322:2377–8.

20 Krause J, Gulshan V, Rahimy E, *et al*. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 2018;125:1264–72.
21 Benson T. *Principles of health Interoperability HL7 and SNOMED*. London, England: Springer, 2012. ISBN: 978-1-4471-2800-7.
22 FitzHenry F, Resnic FS, Robbins SL, *et al*. Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. *Appl Clin Inform* 2015;6:536–47.
23 Rosenbloom ST, Carroll RJ, Warner JL, *et al*. Representing knowledge consistently across health systems. *Yearb Med Inform* 2017;26:139–47.
24 Lehne M, Luijten S, Vom Felde Genannt Imbusch P, *et al*. The use of FHIR in digital health - A review of the scientific literature. *Stud Health Technol Inform* 2019;267:52–8.
25 Bruland P, McGilchrist M, Zapletal E, *et al*. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. *BMC Med Res Methodol* 2016;16:1.
26 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
27 Balki I, Amirabadi A, Levman J, *et al*. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J* 2019;70:344–53.
28 Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;3:1157–82.
29 Moons KGM, de Groot JAH, Bouwmeester W, *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS Med* 2014;11:e1001744.
30 Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. *JAMA* 2019;322:1351–2.
31 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.
32 Coiera E, Ammenwerth E, Georgiou A, *et al*. Does health informatics have a replication crisis? *J Am Med Inform Assoc* 2018;25:963–8.
33 Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A* 2016;113:7345–52.
34 Saria S, Subbaswamy A. Tutorial: safe and reliable machine learning. arXiv.org, 2019. Available: https:// arxiv.org/abs/1904.07204
35 Li X, Zhang S, Zhang Q, *et al*. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol* 2019;20:193–201.
36 Zech JR, Badgeley MA, Liu M, *et al*. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018;15:e1002683.
37 Soleimani H, Hensman J, Saria S. Scalable joint models for reliable Uncertainty-Aware event prediction. *IEEE Trans Pattern Anal Mach Intell* 2018;40:1948–63.
38 Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data* 2016;3:9.
39 Tschandl P, Rinner C, Apalla Z, *et al*. Human-computer collaboration for skin cancer recognition. *Nat Med* 2020;26:1229–34.
40 et alCaruana R, Lou Y, Gehrke J. Intelligible algorithms for healthcare: predicting pneumonia risk and hospital 30-day readmission. *Paper presented at: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*2015.
41 Oakden-Rayner L. *Exploring the ChestXray14 dataset: problems*. Wordpress: Luke Oakden Rayner, 2017.
42 Zhang Z, Beck MW, Winkler DA, *et al*. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Transl Med* 2018;6:216.
43 Nielsen AB, Thorsen-Meyer H-C, Belling K, *et al*. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish national patient registry and electronic patient records. *Lancet Digit Health* 2019;1:e78–89.
44 Steiner DF, MacDonald R, Liu Y, *et al*. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol* 2018;42:1636–46.
45 Kuzniewicz MW, Puopolo KM, Fischer A, *et al*. A quantitative, risk-based approach to the management of neonatal early-onset sepsis. *JAMA Pediatr* 2017;171:365–71.
46 Cronin PR, Greenwald JL, Crevensten GC, *et al*. Development and implementation of a real-time 30-day readmission predictive model. *AMIA Annu Symp Proc* 2014;2014:424–31.
47 Miller A, Koola JD, Matheny ME, *et al*. Application of contextual design methods to inform targeted clinical decision support interventions in sub-specialty care environments. *Int J Med Inform* 2018;117:55–65.
48 Houssami N, Lee CI, Buist DSM, *et al*. Artificial intelligence for breast cancer screening: opportunity or hype? *Breast* 2017;36:31–3.
49 Clifton DA, Niehaus KE, Charlton P, *et al*. Health informatics via machine learning for the clinical management of patients. *Yearb Med Inform* 2015;10:38–43.
50 Giannini HM, Ginestra JC, Chivers C, *et al*. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med* 2019;47:1485–92.
51 Khalifa M, Magrabi F, Gallego B. Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support. *BMC Med Inform Decis Mak* 2019;19:207.
52 Xie Y, Gunasekeran DV, Balaskas K, *et al*. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. *Transl Vis Sci Technol* 2020;9:22.
53 O'Neil C. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. London: Allen Lane, 2016.
54 Palmer A. *IBM's Watson AI suggested "often inaccurate" and "unsafe" treatment recommendations for cancer patients, internal documents show*. DailyMail.com, 2018. https://www.dailymail.co.uk/sciencetech/article-6001141/IBMs-Watson-suggested-inaccurate-unsafe-treatment-recommendations-cancer-patients.html?ito=email_share_article-top
55 Challen R, Denny J, Pitt M. Artificial intelligence. *bias and clinical safety BMJ Qual Saf* 2019;28:231–7.
56 Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle regulation of artificial intelligence– and machine learning–based software devices in medicine. *JAMA* 2019;322:2285–6.
57 Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. *J Am Med Inform Assoc* 2019;26:1645–50.
58 Finlayson SG, Bowers JD, Ito J, *et al*. Adversarial attacks on medical machine learning. *Science* 2019;363:1287–9.
59 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017;24:423–31.
60 Komorowski M, Celi LA. Will artificial intelligence contribute to overuse in healthcare? *Crit Care Med* 2017;45:912–3.
61 Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med* 2018;378:981–3.
62 Carter SM, Rogers W, Win KT, *et al*. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast* 2020;49:25–32.
63 Abràmoff MD, Tobey D, Char DS. Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. *Am J Ophthalmol* 2020;214:134–42.
64 Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA* 2019;322:1765–6.
65 Jiang JX, Bai G. Types of information compromised in breaches of protected health information. *Ann Intern Med* 2020;172:159–60.
66 AI ethics guidelines global inventory. Available: https://inventory.algorithmwatch.org/;
67 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.
68 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.
69 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.