Recommendations for the development and use of imaging test sets to investigate the test performance of artificial intelligence in health screening

Anastasia Chalkidou, Farhad Shokraneh, Goda Kijauskaite, Sian Taylor-Phillips, Steve Halligan, Louise Wilkinson, Ben Glocker, Peter Garrett, Alastair K Denniston, Anne Mackie, Farah Seedat

Rigorous evaluation of artificial intelligence (AI) systems for image classification is essential before deployment into health-care settings, such as screening programmes, so that adoption is effective and safe. A key step in the evaluation process is the external validation of diagnostic performance using a test set of images. We conducted a rapid literature review on methods to develop test sets, published from 2012 to 2020, in English. Using thematic analysis, we mapped themes and coded the principles using the Population, Intervention, and Comparator or Reference standard, Outcome, and Study design framework. A group of screening and AI experts assessed the evidence-based principles for completeness and provided further considerations. From the final 15 principles recommended here, five affect population, one intervention, two comparator, one reference standard, and one both reference standard and comparator. Finally, four are appliable to outcome and one to study design. Principles from the literature were useful to address biases from AI; however, they did not account for screening specific biases, which we now incorporate. The principles set out here should be used to support the development and use of test sets for studies that assess the accuracy of AI within screening programmes, to ensure they are fit for purpose and minimise bias.

Introduction

Many screening programmes, such as breast cancer or diabetic eye screening, use medical images to detect early disease. Human image interpretation is usually subjective, labour-intensive, and resource-intensive. Advances in computing power and algorithm design, and increasing access to large datasets are accelerating the development of artificial intelligence (AI) systems to support image interpretation. Indeed, several AI products claim similar or more accurate diagnostic performance compared with human interpretation.[1,2] Developing AI for use in screening is a multistep process, which includes the evaluation of the accuracy, clinical effect, cost-effectiveness, and ethical implications of AI. This evaluation ensures that decisions to implement AI in screening are supported with robust evidence and that deployment can be safely achieved.

There are three datasets in the development and evaluation of AI: a training set to develop the AI, a tuning set to tune hyperparameters, and a test set to evaluate diagnostic performance.[3] Although development data are often reused to fine-tune the AI (termed internal evaluation), a test set (also termed a validation test) of data or images that have not been used for its development are assessed for external evaluation (termed validation). External evaluation is a crucial early step to evaluate diagnostic performance. External evaluation is usually a retrospective study on previously collected images and is conducted after developing the algorithm but before prospective evaluation in clinical practice when it is deployed in a real-world service setting.[4]

The exact nature of this test set and the study design of this phase of evaluation remains a subject of debate and there is no consensus to support developers or regulators to determine test-set quality.[5,6] Therefore, we aimed to propose a set of principles that could be used when curating test sets and designing studies using these sets to assess the performance of AI for image classification in screening. To achieve this aim, we conducted a rapid review and evidence synthesis of the existing literature on methods to develop test sets and consulted on identified principles with an expert group. Finally, we tested the applicability of the principles by applying them to two studies investigating the accuracy of AI using test sets.[1,2] The focus of this guidance is on test sets only and considering the principles of development datasets is outside its scope.[3]

Principles for developing and using a test set for AI evaluation

We used thematic analysis for the qualitative synthesis, extracting themes from the literature and mapping them to a framework. We used the Population, Intervention, and Comparator or Reference standard, Outcomes, and Study design framework for coding the list of principles, which is the main framework used by the UK's National Institute for Health and Care Excellence, the US Food and Drug Administration, and other major health technology assessment bodies. It is a framework that facilitates comparative effectiveness research and can be used to guide evaluation studies of AI systems.

The list of principle themes found in the literature was shared with the UK National Screening Committee's AI task group. The group reviewed the themes and added screening-specific considerations to reach a final list of principles via consensus through an iterative process. The final list of principles is reported in this paper along with their source.

The key principles for developing a test set for AI evaluation are divided into those found in the literature

and those proposed by the expert group (table). These principles are discussed in more detail here and examples of how they can be applied to breast and diabetic eye screening are discussed in the appendix (pp 11–14).

Population

The test set should be representative of the target population

The test set should represent the target population that the AI will be applied to. The number of included positive and negative cases is crucial and must reflect the intended AI use case.[4,7–12] There is consensus that prevalence has an effect on the performance of AI.[3,4,9–13] This impact can occur mostly by having a prevalence rate in a test set that is not representative of the target population.[14] The test set should include all healthy, benign, and pathological states relevant to the intended use of AI to minimise spectrum bias. Spectrum bias can occur when the test set does not appropriately represent the full range of findings identified in the target population.

Test sets should also include all variations in patient demographics.[10,11] To facilitate understanding of the demographics of the population represented by a test set, the Minimum Information for Medical AI Reporting standards[15] propose the minimum information necessary to generalise findings. The following demographic variables are included: age, sex, race, ethnicity, and socioeconomic status. Specifically, for a national screening programme, the test set should be large enough to represent the full range of risks that the target population could constitute. For example, in breast screening, a test set would need to include people with low risk of cancer, as well as those with a higher risk such as those with dense breasts.

The test set should be independent of the development datasets

The test set should be completely independent of the development dataset to adequately avoid overfitting and avoid data snooping.[3,7,9–12,16–18] Reusing development data for evaluation could overestimate the accuracy in a test-set study, and not reflect real-world performance.

Both temporal and geographical independence are proposed in the literature, although no explicit definitions are provided.[12,19] Broadly, temporal independency refers to the use of data collected in different time periods but from the same institutions that provided development data. Geographical independency refers to data collected from geographically remote institutions.[12] Geographical independency provides a more reliable estimate of AI's generalisability than temporal evaluation as individuals from the same institutions or locations share similar characteristics. Images from different institutions can also capture natural and unavoidable variations in scanner and image acquisition parameters (eg, scanners, staff, protocols, and contrast material). Generalisability from temporal validation studies could be particularly problematic for screening test sets as often the same individuals' old data might be part of the development dataset and their new data part of the test set. Therefore, truly independent datasets should be used to assess generalisability to the population of interest.[10,13,19]

The test set should be multicentred

Test sets should be multicentred because the diversity of data collected and the probable representativeness of the population sampled is increased. A multicentred dataset would enhance the generalisability of results, especially for cases in which the spectrum of disease, population characteristics, or the technical characteristics of imaging could vary according to site.[10,12,16,18,19]

The dataset should consist of images collected consecutively (or randomly)

Collecting data selectively in a case-control manner to ensure a certain ratio between patients who are disease positive and those who are disease negative should be avoided for the test set.[10–12,19] The artificial ratio between two extreme populations of individuals who are disease positive and disease negative introduces spectrum bias and an atypical ungeneralisable disease cohort, that could inflate diagnostic accuracy measures such as sensitivity and specificity. A diagnostic cohort design, in which patients are recruited consecutively or randomly, is recommended for the test set. This method is because the clinical setting and patient eligibility criteria are prospectively defined on the basis of the target population (eg, those who attend screening) and not around their screening or diagnostic outcome. Consequently, the test set is more likely to represent the target population and spectrum of disease in clinical practice.

However, consecutive data collection might not allow rare pathologies to be captured unless very large samples are used, which might not be feasible. Therefore, enrichment of positive cases might be necessary. To do this enrichment with minimal bias, hybrid studies could be useful. Hybrid studies are better than case-control studies as they allow all consecutively enrolled individuals with the disease, and a randomly selected sample of consecutively enrolled individuals without the disease, to be included (similar to a nested case-control study). This method does mean, however, that the range of individuals without disease to inform specificity is not as wide as in a pure consecutive sample.[20]

The test set should account for technical variations in image acquisition, including image quality

Technical image acquisition parameters and quality differ across centres and patients. When testing AI performance, including images that represent these variations so that their effect on accuracy can be assessed is important. For example, breast screening test sets should include films of mixed compression, exposure factors, filters, positioning, technical repeats, and number or types of views.

	Identified in the literature	Further considerations proposed by the UK National Screening Committee Artificial Intelligence task group
Population	The test set should represent the whole spectrum of pathological and normal findings encountered in the target population as well as the key demographics	The dataset should be representative of the real screening population, including the full age and ethnic diversity of the UK population; it should be sufficiently large to represent women with varying levels of risk and have uncommon events such as rare breast pathologies and varied mammographic features
Population	The test set should be independent of the development datasets	Specific to screening, generalisability from temporal validation studies could be problematic for test set studies in screening settings as often the same individuals' old data might be part of the development dataset and their new data part of the test set
Population	The test set should be multicentred	No further comment
Population	The dataset should consist of images collected consecutively (or randomly)	However, consecutive data collection might not allow rare pathologies to be captured unless very large samples are used, which might not be feasible; random data collection could be helpful in this case; for example, hybrid studies could be used when all consecutively enrolled individuals with disease are included and a randomly selected sample of the consecutively enrolled individuals without disease (similar to a nested case-control study); this method is likely to ensure that rare pathologies are captured
Population	The test set should account for technical variations in image acquisition, including image quality	For breast cancer screening, the test set should include films of mixed technical quality (eg, compression, exposure factors, filters, and positioning; including technical repeats, and number and types of views); when AI is proposed as the first reader of multiple readers in a screening programme, the threshold of technical recalls due to an inability to process the data for AI scrutiny can then be compared with the existing rates of technical recalls for that programme; with respect to image quality, there could be a systemic issue in the use of retrospective test sets if they are only taken from the final set of images from clinical practice; knowing how many times the image was taken (ie, a clinician could not read the image, so it was re-taken until it could be read) could be difficult; this issue should be taken into account when test sets are being considered
Intervention	A particular test set should only be used a limited number of times on different versions of the same AI system and repeated testing should be explicit	No further comment
Comparator	The level of expertise in the comparator should be similar to the standard of care	In the study, the level of expertise of readers in the comparator group should be compared against the standard of care and the comparator reading should take place in a clinical practice; national screening programmes have pre-specified requirements for the training and performance requirements for human reader grading and reporting
Comparator	Not identified	The comparator reading should take place in clinical practice
Reference standard	Mislabelling should be minimised (ie, misclassification)	The choice of an appropriate reference standard to avoid mislabelling will also depend on its intended clinical pathway (eg, replacing a human reader, triage, or add on); screening programmes aim to detect disease early and are subject to additional sources of bias that can affect the choice of a reference standard such as lead time bias, length bias, differential verification bias, and overdiagnosis
Reference standard or comparator	Interobserver agreement should be reported	No further comment
Outcome	The analysis of the test set should report relevant outcomes	The analysis should report test accuracy including sensitivity and specificity, and positive and negative predictive value at UK screening prevalence; the area under the receiver operating curve is useful for deciding the threshold during the training and tuning steps; it is less useful in screening than these outcomes as it fails to consider results at diagnostically important thresholds, which is pivotal in screening when the decision is binary (ie, to recall or not)
Outcome	The analysis of the test set should report the threshold at which accuracy estimates are reported	The choice of the threshold should not be arbitrary but should be determined by the way the AI system will be used in the pathway (eg, to maximise sensitivity, specificity, or both); for example, if AI in diabetic eye screening is used as a pre-screening tool or replacement of level one graders, maximising sensitivity is more important
Outcome	The choice of outcome measures should consider the presence of class imbalance	No further comment
Outcome	Evaluation outcomes should be reported with CIs	No further comment
Study design	A formal sample size calculation should be shown	The test set should be large enough to include uncommon events such as rare breast pathologies and varied mammographic features and be powered to identify artificial intelligence capability in these individuals

Table: List of principles identified in the literature and those proposed by the UK National Screening Committee Artificial Intelligence task group

Evaluation results should include the proportion of individuals for whom AI did not analyse images because of technical reasons such as poor image quality.[8] When AI is proposed as the first of multiple readers in a screening programme, the threshold of technical recalls can then be compared with the existing rates of technical recalls for that programme. For example, the UK breast cancer screening programme standards have a less than 0·7% acceptable threshold for technical recalls and less than 2·0% for technical repeats whereas the data encryption standards have a 2·0–4·0% acceptable level of ungradable images.[21,22] However, the use of retrospective test sets could be problematic in this regard if only the final set of images from clinical practice is recorded; it might be difficult to know how many times the image was re-taken until it could be read by a clinician.

Intervention or index test

A particular test set should only be used a limited number of times on different versions of the same AI system and repeated testing should be explicit

Running multiple versions of the same AI system with different hyperparameters, then testing them all against the same test set, and then retrospectively choosing the parameters that achieve the highest diagnostic accuracy increases the likelihood that accuracy is inflated

compared with daily practice. This testing of different versions is similar to the tuning step and to running multiple statistical tests on the same dataset and selectively reporting those that achieve statistical significance (ie, p-hacking bias). In effect the dataset is being used for hypothesis generation rather than hypothesis testing. Only a limited number of versions of one AI system should ever be tested against the same test set, and the criteria for selecting which models will be tested should be explicit a priori.[10]

Comparator
Level of expertise
The comparator in a test-set study should be the standard of care and will often involve test interpretation by a qualified individual. Authors should report the level of expertise for each reader (eg, subspecialty training, years of clinical experience, and volume of individuals positive for disease reported each year).[10,18] The level of expertise required should be aligned to the standard of care to avoid straw man bias, in which the performance of an AI algorithm is compared against human reviewers who do not have representative expertise.[10] Often national screening programmes have pre-specified requirements for the training and performance of human interpretation.[23]

The comparator reading should take place in clinical practice
Human reading should take place in clinical practice and not in a laboratory setting, as the latter can introduce bias due to the laboratory effect.

Reference standard—outcome
Minimise mislabelling (ie, misclassification)
Methods of labelling (ie, classifying medical images) include using historical radiology reports, expert consensus, reference standard imaging or laboratory examinations, clinical outcomes, and surgical or pathological confirmation. Although some methods of labelling are better than others (ie, the gold standard methods), they are not always feasible or achievable. For example, gold standard labelling of breast cancer in mammography includes biopsy in breast screening but, as an invasive procedure, it is not appropriate for individuals who are screen negative. Unavailability of the gold standard can lead to mislabelling in the test set, which can be a serious issue.[10] Misclassification might also arise from human readers. Humans are fallible, and therefore defining the gold standard or ground truth by their subjective judgement is problematic.[16] AI that in reality outperforms a human reader will appear inferior because it will disagree with an imperfect reference standard.[9] This situation is unavoidable in some diseases, such as diabetic retinopathy, as the reference standard is always human (eg, a grader or ophthalmologist).[14] In such cases, readers should be adequately qualified and, when possible, an expert consensus panel will be the preferred choice.[8]

The choice of an appropriate reference standard to avoid mislabelling depends on the use case of AI in the clinical pathway, and screening pathways, in particular, are subject to sources of bias such as verification bias and over-diagnosis, that can affect the choice of a reference standard. For example, in the case of breast cancer screening, further testing, such as biopsy, or long-term follow-up of interval cancers can be used as the reference standard. If the role of AI in screening is to find cancer missed by the human reader (ie, as an add on), then using the interval cancer rate as the reference standard would achieve more generalisable and valid results than a human reader. However, when AI is aiming to replace a human reader, the validity of a test set depends on whether individuals who are positive are limited to those who are detected by an imperfect human reader, or whether they also include individuals who were diagnosed with cancer but were reported as negative by the human reader. In this setting, the test set will need to use biopsy for individuals who are screen positive and detection of interval cancers from long-term follow-up for individuals who are screen negative (ie, differential verification). Including the long-term follow-up will then minimise verification bias, although it will not completely remove it.[20]

Reference standard or comparator
Interobserver agreement should be reported
Interobserver agreement among readers should always be reported as it is an important feature of the reliability of observer opinion, especially when used as the reference standard.[4,10,16] The use of maps that visualise areas of uncertainty has been proposed to help reflect variability of expert opinion regarding the same images.[16] Applying uncertainty maps could have implications on how the dataset is collected, requiring in some cases prospective data collection. Cases with low interobserver agreement should not be removed as spectrum bias could be introduced.

Outcome measures
The analysis of the test set should show relevant outcome measures
Using the outcome measure of number of correct predictions versus total number of predictions can be misleading as it depends on disease prevalence.[10] For example, a binary classifier that diagnosed a rare condition present in only 0·1% of individuals would achieve 99·9% accuracy simply by calling all cases negative. Instead, outcome measures such as positive and negative predictive values at appropriate prevalence rates, sensitivity, and specificity are recommended. Studies should also show accuracy for clinically important pathology subtypes and by relevant participant characteristics, for example, age and ethnicity.

The area under the receiver operating characteristic curve is a frequently reported performance metric that combines sensitivity and specificity in a single metric and is often requested by regulatory bodies such as the US Food and

Drug Administration. The use of this performance metric should reflect the context of the development stage. The area under the receiver operating characteristic curve is useful to determine the positivity threshold during training and tuning steps but might not consider results at diagnostically important thresholds, which is pivotal in screening when the decision is binary (ie, to recall or not) and when the consequences of false-positive versus false-negative decisions are not equivalent. For example, if AI in diabetic eye screening is used as a pre-screening tool or replacement of level one graders, maximising sensitivity is more important than maximising specificity. The area under the curve is therefore of less value than providing sensitivity, specificity, and positive and negative predictive values, with supporting contingency tables.

The test-set analysis must show the thresholds at which accuracy estimates are reported

As a minimum, contingency tables (ie, including true positives, true negatives, false positives, and false negatives), should be shown at justified pre-specified thresholds.[3,4,18,24] The choice of the threshold should not be arbitrary but should depend on the use case of AI (for example, if AI is used in the data encryption standard as a pre-screening tool or as level one graders, a threshold that maximises sensitivity is required).

The choice of outcome measures should take into consideration the presence of class imbalance

Class imbalance arises when disease categories or classes are not represented equally in the dataset, which is common in screening programmes. Again sensitivity, specificity, and negative and positive predictive values are the recommended outcomes in this context, whereas the area under the curve will overestimate AI performance.

Precision, recall, F1 measures (ie, an overall measure of a model's accuracy that combines precision and recall), and area under the precision-recall curve have also been mentioned in the literature for use in the presence of class imbalance.[16,25] However, the F1 score assigns an equal cost to false-negative and false-positive results, which is not often true for screening, as mentioned previously.[10] The area under the precision-recall curve shows the trade-off between precision and recall across different decision thresholds and is suggested as a better metric than the area under the curve.[24]

Evaluation outcomes should be shown with CIs

A confidence interval should be shown for the outcome measures, for example 95%, to adequately capture uncertainty around the results' generalisability to the population.[10,13,16]

Study design—statistics

A formal sample size calculation should be shown

The sample size should be shown and justified.[3,12,18,19,25] As a minimum, the test-set sample size will be determined according to the study hypothesis (eg, equivalence, non-inferiority, or superiority), the intended use (eg, binary classification *vs* diagnosis of multiple outcomes), and the minimal difference considered clinically acceptable. For example, in breast cancer screening, the test set should be large enough to include uncommon events such as rare breast pathologies and varied mammographic features and be powered to identify AI capability in these cases.

Discussion

Well designed test sets are key to provide unbiased evaluations of an AI system's diagnostic accuracy after training and tuning. This guidance is the first of its kind to identify a set of principles for the development and use of test sets that are fit for purpose when assessing the diagnostic accuracy of AI image classification in screening. Although the proposed principles from the literature aimed at addressing the main sources of bias for AI image classification (such as spectrum bias, overfitting, straw man bias, data snooping bias, and p-hacking bias), they did not account for the intricacies of screening. As the purpose of screening is early and presymptomatic detection, additional biases such as verification bias and overdiagnosis require consideration in developing test sets. By proposing these additional considerations, we have not only reviewed, but also furthered, the discussion in this area. The identified principles should improve the validity and generalisability of estimates of AI performance in imaging-based screening programmes.

Applying the proposed principles might be challenging in some cases because of feasibility concerns. For example, although test sets should be representative of the target population,[15,26] as emphasised by the concept of Health Data Poverty (the importance of having representative data when developing data-driven technologies),[27] existing studies often do not adequately describe such parameters.[8,15] Hard to reach populations often have limited engagement with research or health-care services and might be under-represented in datasets. Additionally, access to the required data, such as ethnicity and socioeconomic status, could be considered sensitive information and hindered by information governance restrictions. This difficulty could leave some groups of people unable to benefit, or even harmed, from AI further widening the gap in health inequalities. A potential solution when patients' confidential data is not accessible for research, is for data providers to provide summary statistics. When the research team does have access to confidential information, that cannot be made publicly available, researchers can report summary statistics at the population level. For example, the screening attendance rate by deprivation level for a consecutively included population can be reported as a surrogate for socioeconomic status.

Several included papers discussed the issue of sample size calculation and the need to establish an appropriate hypothesis and choose a clinically significant effect size.[3,9,12,19,25] In previous systematic reviews of AI in medical

Search strategy and selection criteria

We followed the rapid review recommendations proposed by the Cochrane Rapid Reviews Methods Group. The search strategy combined MeSH and free text terms on three themes: technology (artificial intelligence), context (medical imaging), and study design. We searched relevant literature databases for papers published from Jan 1, 2012 to Sept 11, 2020. The search was supplemented with forward snowballing of the references of included studies.

Eligible studies were systematic reviews (if the authors included methodological quality in their analysis), methodology papers, critical appraisal tools and supporting documents, reporting guidelines, and opinion pieces (if they were representing views of, for example, a radiological society rather than individuals) that recommend standards or principles for the development of test sets used for the evaluation of (artificial intelligence) AI systems. The setting could be diagnosis or screening and could include any medical imaging modality. Included papers had to list at least one principle. Papers that referenced principles published elsewhere in the discussion section were not included. There were no restrictions on the underlying AI system or medical condition.

18 full text articles out of 3441 screened abstracts were included; five systematic reviews, two systematic reviews and meta-analyses, three narrative reviews, six guidance documents, one perspective, and one reporting standard. 12 papers referred generally to AI systems, two referred to a convolutional neural network, three to deep learning and two to machine learning. Four papers had screening as the clinical setting, whereas the majority referred to the field of medical imaging more broadly. Generally, the papers were of moderate quality. Details on the included studies are provided in the appendix (p 5) as are more details about the search and selection process, quality appraisal, data extraction, and synthesis (appendix p2). The protocol is registered on Open Science Framework.

imaging, only two of 171 papers provided a sample size estimation.[3,19] The clinically significant effect size will vary depending on the use case or outcome and reference standard. For example, although not AI-related, the PROSPECTS trial (NCT03733106), that investigates the role of tomosynthesis in the UK Breast Cancer Screening Programme, estimated that a sample size of 100 000 women attending screening will be required to show a difference of 1 of 1000 women with interval cancers.

Publicly available datasets for the development of AI systems exist; however, the suitability of these datasets for the evaluation of AI is unknown. A review of AI for diabetic retinopathy screening showed that Messidor-2 and Kaggle-DR were used as test sets in eight (ie, four each) of the 11 listed studies.[8] Messidor-2 includes images of relatively high quality, and only 4% are deemed of insufficient image quality, which might not be a good representation of data from screening programmes.[28] For breast cancer screening, the UK-based Optimam and the Swedish Cohort of Screen-Aged Women databases include a large number of mammograms representative of their respective national screening programmes.[29,30] In both databases, screening decisions and clinical outcome data were also collected by linkage. The list of principles reported here can be used to appraise such datasets and improve their quality.

Present methodological limitations for the development of test sets could in the future be addressed by technical advances. For example, methods such as transfer learning, a machine learning method in which a model developed for a task is reused as the starting point for a model on a second task, could help optimise an AI system for a new target population.[31] The use of occlusion testing could help improve our ability to explain misclassifications and weighted-error scoring could help assess the effect of those misclassifications.[32] Finally, the use of deep learning algorithms for automated labelling across new image modalities could help replace the use of human expert consensus for defining a gold standard with objective information obtained from a different imaging modality.[33]

Strengths of this paper include that it is the first paper summarising principles of developing and using test sets that are generalisable across medical imaging and that it provides further considerations that reflect the specifics of screening programmes; we used a systematic search and experts in the field; and the guidance is presented using the Population, Intervention, and Comparator or Reference standard, Outcomes, and Study design framework facilitating its use by health technology assessment bodies. Some limitations should be noted. The search was limited to 2012–20, potentially missing relevant publications. However, this time limitation is aligned with other systematic reviews in this field.[3] Publications from 2021 to 2022, while the paper was undergoing peer review, were not included. We also complemented our search with snowballing without date limits thus minimising the risk of missing publications. Only 20% of the search results and data extractions were double checked. Although this could have potentially increased the possibility of error, the nature of the review focusing on qualitative synthesis rather than quantitative makes the effect of such error less likely.

Conclusion

The list of principles set out in this paper should help developers and researchers curate high-quality test sets, suitable for pre-deployment evaluation. Health service gatekeepers such as regulators, health technology assessment bodies, and screening service providers can also apply the principles set out here to assess the quality of test sets and the quality of evaluation conducted in such studies, providing them with greater confidence. As this field will evolve and we will learn from the experience of evaluating different AI algorithms, our understanding of the principles will also evolve, and these recommendations can be further refined.

References
1   McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577:** 89–94.
2   Heydon P, Egan C, Bolter L, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *Br J Ophthalmol* 2021; **105:** 723–28.
3   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.
4   Faes L, Liu X, Wagner SK, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. *Transl Vis Sci Technol* 2020; **9:** 7.
5   Lee CI, Houssami N, Elmore JG, Buist DSM. Pathways to breast cancer screening artificial intelligence algorithm validation. *Breast* 2020; **52:** 146–49.
6   Wang S, Zhang Y, Lei S, et al. Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: a systematic review and meta-analysis of diagnostic test accuracy. *Eur J Endocrinol* 2020; **183:** 41–49.
7   Brinker TJ, Hekler A, Utikal JS, et al. Skin cancer classification using convolutional neural networks: systematic review. *J Med Internet Res* 2018; **20:** e11936.
8   Nielsen KB, Lautrup ML, Andersen JKH, Savarimuthu TR, Grauslund J. Deep learning-based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance. *Ophthalmol Retina* 2019; **3:** 294–304.
9   Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl Vis Sci Technol* 2020; **9:** 42.
10 England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. *AJR Am J Roentgenol* 2019; **212:** 513–19.
11 Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol* 2019; **20:** 405–10.
12 Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018; **286:** 800–09.
13 Murtagh P, Greene G, O'Brien C. Current applications of machine learning in the screening and diagnosis of glaucoma: a systematic review and meta-analysis. *Int J Ophthalmol* 2020; **13:** 149–62.
14 Yip MYT, Lim G, Lim ZW, et al. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. *NPJ Digit Med* 2020; **3:** 40.
15 Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc* 2020; **27:** 2011–15.
16 Sanchez-Peralta LF, Bote-Curiel L, Picón A, Sanchez-Margallo FM, Pagador JB. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. *Artif Intell Med* 2020; **108:** 101923.
17 Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The algorithmic audit: working with vendors to validate radiology-AI algorithms—how we do it. *Acad Radiol* 2020; **27:** 132–35.
18 Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the *Radiology* Editorial Board. *Radiology* 2020; **294:** 487–89.
19 Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. *Eur J Nucl Med Mol Imaging* 2019; **46:** 2656–72.
20 Taylor-Phillips S, Seedat F, Kijauskaite G, et al. UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. *Lancet Digit Health* 2022; **4:** e558–65.
21 Public Health England. NHS breast screening programme screening standards valid for data collected from 1 April 2021. 2021. https://www.gov.uk/government/publications/breast-screening-consolidated-programme-standards/nhs-breast-screening-programme-screening-standards-valid-for-data-collected-from-1-april-2021 (accessed June 1, 2021).
22 Public Health England. Diabetic eye screening standards valid for data collected from 1 April 2019. March 12, 2021. https://www.gov.uk/government/publications/diabetic-eye-screening-programme-standards/diabetic-eye-screening-standards-valid-for-data-collected-from-1-april-2019 (accessed June 1, 2021).
23 Public Health England. Public health functions to be exercised by the NHS Commissioning Board Service specification; no.22 NHS diabetic eye screening programme service specification. 2012. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/213169/22-NHS-Diabetic-Eye-Screening-Programme-Service-Specification.pdf (accessed June 1, 2021).
24 Yanagihara RT, Lee CS, Ting DSW, Lee AY. Methodological challenges of deep learning in optical coherence tomography for retinal diseases: a review. *Transl Vis Sci Technol* 2020; **9:** 11.
25 Koçak B, Durmaz E, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol* 2019; **25:** 485–95.
26 Massat NJ, Douglas E, Waller J, Wardle J, Duffy SW. Variation in cervical and breast cancer screening coverage in England: a cross-sectional analysis to characterise districts with atypical behaviour. *BMJ Open* 2015; **5:** e007735.
27 Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021; **3:** e260–65.
28 Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci* 2016; **57:** 5200–06.
29 Halling-Brown MD, Warren LM, Ward D, et al. OPTIMAM mammography image database: a large-scale resource of mammography images and clinical data. *Radiol Artif Intell* 2020; **3:** e200103.
30 Dembrower K, Lindholm P, Strand F. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the Cohort of Screen-Aged Women (CSAW). *J Digit Imaging* 2020; **33:** 408–13.
31 Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health* 2019; **1:** e35–44.
32 Gunasekeran DV, Wong TY. Artificial intelligence in ophthalmology in 2020: a technology on the cusp for translation and implementation. *Asia Pac J Ophthalmol (Phila)* 2020; **9:** 61–66.
33 Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology* 2019; **126:** 513–21.