Contemporary Clinical Trials Communications 16 (2019) 100434

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Research paper

Statistical considerations for testing an AI algorithm used for prescreening lung CT images

Nancy A. Obuchowski*, Jennifer A. Bullen

Quantitative Health Sciences /JJN3, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH, 44195, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Diagnostic accuracy
Prescreening
Computer-aided detection
Diagnostic accuracy studies
Area under the ROC curve

ABSTRACT

Artificial intelligence, as applied to medical images to detect, rule out, diagnose, and stage disease, has seen enormous growth over the last few years. There are multiple use cases of AI algorithms in medical imaging: first-reader (or concurrent) mode, second-reader mode, triage mode, and more recently prescreening mode as when an AI algorithm is applied to the worklist of images to identify obvious negative cases so that human readers do not need to review them and can focus on interpreting the remaining cases. In this paper we describe the statistical considerations for designing a study to test a new AI prescreening algorithm for identifying normal lung cancer screening CTs. We contrast agreement vs. accuracy studies, and retrospective vs. prospective designs. We evaluate various test performance metrics with respect to their sensitivity to changes in the AI algorithm's performance, as well as to shifts in reader behavior to a revised worklist. We consider sample size requirements for testing the AI prescreening algorithm.

1. Introduction

Artificial Intelligence (AI) is being applied to medical radiographic images for the purpose of detecting, ruling out, diagnosing, and staging disease. AI algorithms can learn features and then use the insights to assist clinical practice by reducing diagnostic errors and providing outcome prediction; this can all be done in a short period of time, relative to human reader interpretation [1]. Diagnostic imaging has seen the most AI applications in medicine to date, with primary focus on detecting and diagnosing neoplasms [1].

There are four general use cases for AI algorithms in medical imaging: (1) *first*, or *concurrent*, *reader mode*, as when an AI algorithm is applied first to an image and any findings are then reviewed by a human reader during the reader's interpretation, (2) *second-reader mode*, as when an AI algorithm is applied after a human reader has performed his/her interpretation for the reader to consider for inclusion of additional findings detected by the AI, (3) *triage mode*, as when the AI algorithm sorts cases according to suspiciousness of findings, and (4) *prescreening mode*, as when an AI algorithm is applied to a set of images to identify negative cases and then generate the clinical report for this subset of cases, then the human readers focus on interpreting the remaining cases. Computer-aided detection (CAD) with AI algorithms are common in breast, lung, and colon cancer imaging. In these applications, CAD is most commonly used in a second-reader mode, but first-reader and triage modes are gaining acceptance. Prescreening AI algorithms are a relatively new application. They differ from CAD algorithms in that CAD algorithms help readers find lesions but the human readers still interpret all cases and create the reports. In a prescreen AI algorithm use case, human readers never see a subset of cases, and the remaining list of cases differs in prevalence of disease and difficulty. Fig. 1 illustrates these four use cases.

There are three general steps in developing AI tools for medical applications: training, validation (i.e. fine-tuning the model), and testing [2]. The last step, testing, can be performed at three levels: diagnostic performance, patient outcome, and societal efficacy [3]. While there is considerable literature on designing and analyzing studies for testing the diagnostic performance of CAD algorithms in first- and second-reader modes, there is little in the literature about testing prescreening algorithms. In this paper we focus on the statistical considerations for designing a study to test the diagnostic performance a new AI algorithm to identify negative screening chest Computed Tomography (CT) images. We discuss study design, performance metrics, and sample size considerations for this prescreening use case and compare it to CAD studies. We will focus on first- and second-reader modes, as these are currently the most common.

[*] Corresponding author.
E-mail address: obuchon@ccf.org (N.A. Obuchowski).

https://doi.org/10.1016/j.conctc.2019.100434
Received 8 May 2019; Received in revised form 9 August 2019; Accepted 19 August 2019
Available online 22 August 2019

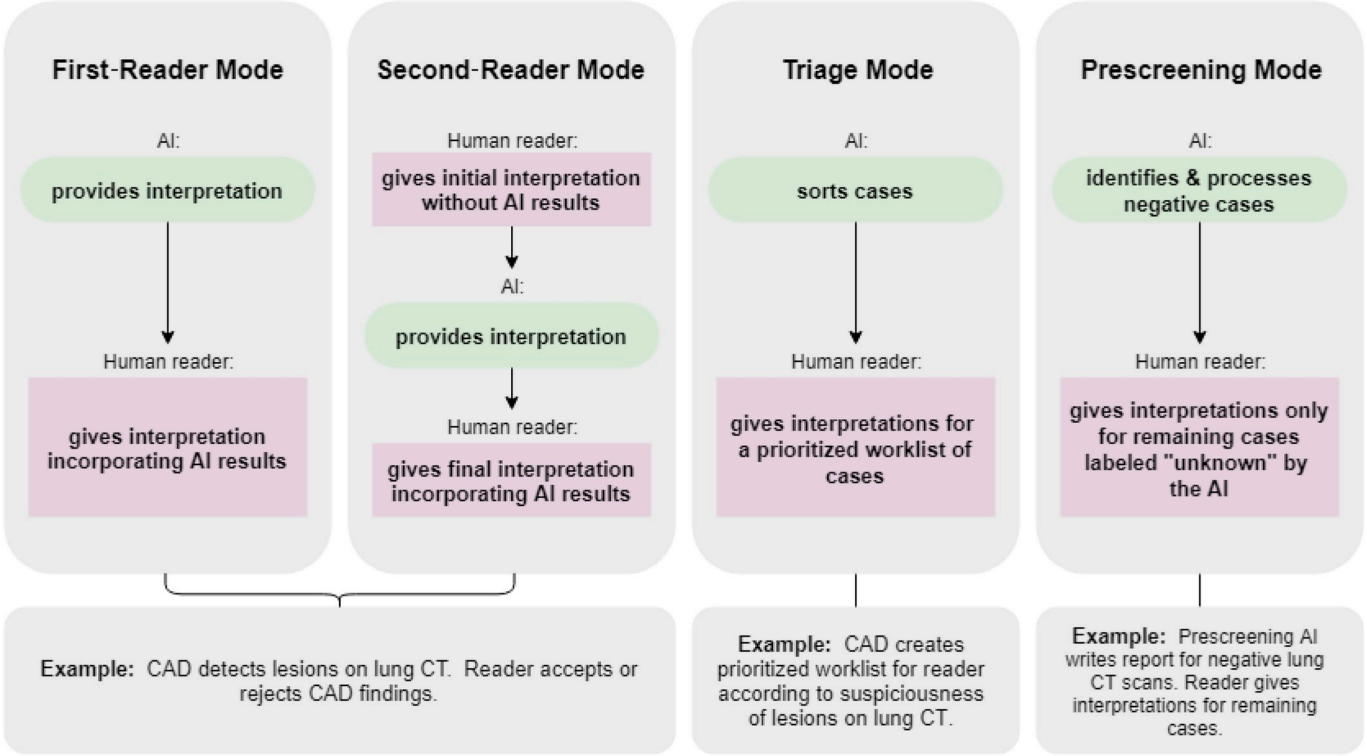
First-Reader Mode
AI:
provides interpretation
Human reader:
gives interpretation incorporating AI results
Second-Reader Mode
Human reader:
gives initial interpretation without AI results
AI:
provides interpretation
Human reader:
gives final interpretation incorporating AI results
Triage Mode
AI:
sorts cases
Human reader:
gives interpretations for a prioritized worklist of cases
Prescreening Mode
AI:
identifies & processes negative cases
Human reader:
gives interpretations only for remaining cases labeled "unknown" by the AI
Example: CAD detects lesions on lung CT. Reader accepts or rejects CAD findings.
Example: CAD creates prioritized worklist for reader according to suspiciousness of lesions on lung CT.
Example: Prescreening AI writes report for negative lung CT scans. Reader gives interpretations for remaining cases.

Fig. 1. Illustration of four use cases for AI

2. Prescreening AI algorithm for detection of lung nodules

Consider a new AI algorithm for lung cancer screening CTs. The AI algorithm acts as a prescreen, classifying cases as either "negative" (i.e. negative for lung cancer) or "unknown". The algorithm generates the report for cases it identifies as "negative" and no further evaluation of the image is performed. Contents of this report are determined by the medical team that uses the AI algorithm The "unknown" cases are interpreted by human readers. The cases that the AI algorithm classifies as "negative" (no further interpretation needed) will predominantly be cases which would be easy for a reader to interpret as negative. It is, of course, possible though for the AI algorithm to classify a case as "negative" that the reader would have classified positive. The cases that the AI algorithm classifies as "unknown" (human reader interpretation needed), will be predominantly be composed of (1) cases that would be a bit harder for a reader to interpret as negative and (2) cases a reader would interpret as positive.

In a study to test the AI algorithm, the performance of human readers with the AI algorithm acting as a prescreen (prescreen arm) would be compared with the readers' diagnostic performance without the AI algorithm (control arm). Fig. 2 illustrates the flow of image interpretations in a test study's control and prescreen arms. The performance of the control arm is fully based on human readers' subjective interpretations, while the performance of the prescreen arm is a hybrid of AI's performance and human readers' interpretations. As with CAD studies, a paired-reader, paired-reader design is typical. This is discussed in more detail in the Study Design section.

The intent of the prescreen AI algorithm is not to improve readers' performance by helping them find lesions, as with a CAD algorithm. Rather, the goal is to reduce the burden on human readers by removing the clearly normal cases, allowing the human readers to concentrate on the more difficult cases. In doing so, the diagnostic performance in the control arm must be maintained in the prescreen arm. Based on a large lung cancer screening study [4], the sensitivity and specificity of human readers interpreting low-dose CT are 0.938 and 0.734, respectively. Thus, a study must be designed to test that the diagnostic performance of readers while using the AI algorithm as a prescreen is not degraded from these performance standards. Simultaneously, the device must identify enough normal cases to be worthwhile.

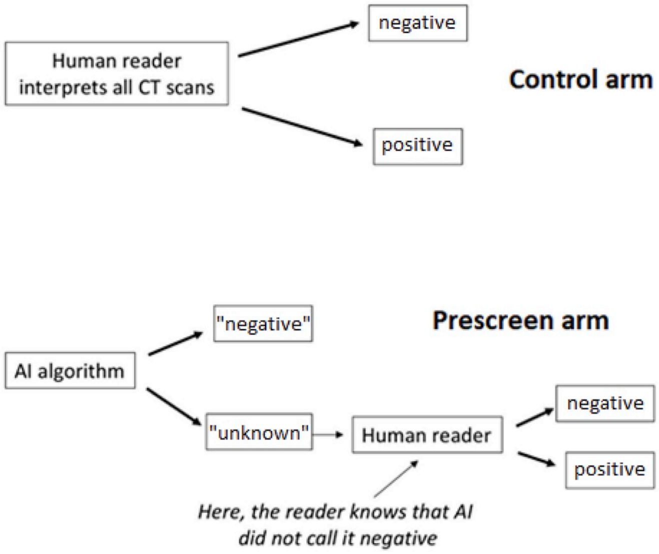

Fig. 2. Illustration of the sequence of interpretations in control and prescreen arms for a prospective study of a prescreening AI algorithm.

We now discuss issues related to designing this study.

3. Study design

3.1. Agreement studies

Agreement studies are common for evaluating AI algorithms because they are easier to conduct and analyze than accuracy studies. In these studies, the interpretations of human readers without the AI are

Table 1
Potential results from agreement study of prescreen AI[a].

Reader finding	Prescreen AI finding	Endpoint Result
negative	"negative"	concordant
negative	"unknown"	concordant
positive	"negative"	discordant
positive	"unknown"	concordant

[a] In an agreement study of a prescreen AI algorithm, the grayed row shows the only situation of discordance.

assessed for concordance/discordance with the interpretations if the prescreen AI had been used. If there is high agreement, then it is concluded that the prescreening with AI can be used without detrimental effect. Agreement studies are often retrospective, making use of existing reader interpretations and comparing them to the results from the AI algorithm.

Table 1 illustrates the potential results for a retrospective agreement study where existing reader interpretations are compared to the results from the AI algorithm. For a prescreening AI algorithm, only cases marked positive by the human reader and "negative" by the AI are discordant (shaded gray in Table 1). Cases interpreted as negative by the human reader are of little value because the prescreen arm and control arm will always be concordant.

Now consider the cases interpreted as positive by the human readers. The agreement rate for these cases (assuming independence in interpretations between human readers and the AI algorithm) can be written:

$$\text{Agreement Rate}_{+} = \frac{\text{Prev}(\text{Sens}_{\text{readers}})(\text{Sens}_{\text{AI}}) + (1-\text{Prev})(1-\text{Spec}_{\text{Readers}})(1-\text{Spec}_{\text{AI}})}{(\text{Sens}_{\text{readers}})\text{Prev} + (1-\text{Spec}_{\text{readers}})(1-\text{Prev})}. \tag{1}$$

where Prev is the prevalence of cancer in the study sample, $\text{Sens}_{\text{readers}}$ and $\text{Spec}_{\text{readers}}$ are the sensitivity and specificity of the human readers in their standard of care interpretation mode, Sens_{AI} is the probability that the AI algorithm determines a cancer case to be "unknown" (aka, the standalone sensitivity of the AI algorithm), and Spec_{AI} is the probability that the AI algorithm determines a non-cancer case to be "negative" (aka, the standalone specificity of the AI algorithm). Note that as the AI's sensitivity increases, the agreement rate increases; in contrast, as the AI's specificity increases, the agreement rate decreases. In other words, as the prescreen AI algorithm improves in specificity, the agreement is worse, suggesting that the prescreen AI is not effective. This is a strong disadvantage of an agreement study for a prescreen AI algorithm, and we do not consider this option further in designing the prescreening study of lung CTs.

3.2. Accuracy studies

An accuracy study is more complicated to design and analyze than an agreement study because all cases need a reference standard to determine the true disease status of the patient. One possible option to reduce the burden of determining truth for all cases might be estimation of *relative* accuracy of the prescreen to control arm [5], but here we consider the situation where truth is known for all cases.

Both retrospective and prospective accuracy studies are possible. In a retrospective design, the clinical site's original readers' interpretations are used in the analysis, whereas in a prospective design, different readers interpret the images in a controlled experimental setting. A prospective design allows investigators to control how and what the readers provide in terms of interpretation. For example, investigators can control whether prior images and/or clinical history is given to the readers, and whether confidence scores (needed for ROC analysis) are collected, in addition to common binary clinical decisions. Furthermore, in testing a prescreen AI algorithm, a prospective study allows investigation of how the prescreening is affecting the readers. Recall from Fig. 2 that when readers interpret after the AI prescreen, their worklist is reduced to the cases classified as "unknown" by the AI algorithm. Reader behavior after the AI prescreen could be affected by: (1) a change in prevalence of disease in the reader's worklist (e.g. if the prevalence of disease is 1%, and if 50% of the normal cases are removed by the AI device, then the prevalence of disease in the worklist nearly doubles), (2) a change in the mix of cases in the reader's worklist (e.g. there is a higher proportion of more difficult normal cases in the worklist), and (3) a subjective reaction to the case being labeled "unknown" by the AI device. These potential effects can be studied in prospective designs but not in a retrospective design. Thus, for testing the prescreen AI algorithm of CT lung images, we consider an accuracy study with prospective reading of the images.

3.3. Comparison with CAD studies

In studies testing CAD algorithms, accuracy studies are the norm because it's critically important to determine if new findings detected with CAD are TPs (i.e. also positive by a reference standard) or FPs (i.e. negative by a reference standard); such a distinction is not possible with an agreement study. Similarly, prospective reader studies are a requirement for testing a CAD algorithm because the human readers interact directly with the CAD findings, accepting or dismissing them. This interaction can only be evaluated in a prospective reader design. In contrast, the prescreen AI algorithm impacts the reader only through changes in the worklist of the remaining "unknown" cases. Research is needed to investigate if and how changes to the worklist impact readers' accuracy. If the impact is negligible, then prospective reader studies may not be needed for testing a prescreen AI algorithm. However, as a first application, a prospective reader design seems necessary for the prescreen lung cancer study.

CAD reader studies often use an enriched study sample with up to 50% of subjects with the disease, based on the reference standard [6,7]]. Similarly, for the lung cancer study using biopsy results and/or two-year follow-up imaging as the reference standard, 50% of the study sample will be subjects with lung cancer and 50% without lung cancer.

The typical CAD study uses a paired-reader paired-case design (i.e. all study readers interpret all study cases in both the control and study arms). A typical reading order randomization scheme for a CAD paired-reader paired-case study can be adopted here for testing the prescreen AI algorithm. The modifications needed for a prescreen AI algorithm are illustrated in Table 2, where a pre-study session is needed for the AI algorithm to be run to identify the "unknown" cases and block 2 readings include only the subset of cases classified by the algorithm as "unknown" instead of all cases. As with CAD studies, readers should receive sufficient training prior to Session 1.

For the subset of cases classified as "unknown" by the AI algorithm, the readers provide an interpretation both without knowledge of the AI result (in Block 1) and with knowledge of the AI result (in Block 2). This subset of cases can be used to investigate the potential impact of the prescreening AI on reader behavior.

4. Performance metrics

Here we focus on discrimination performance indices, rather than calibration performance indices, due to the classification nature of our specific AI example [2]. For a prescreen AI algorithm identifying negative cases as illustrated in Fig. 2, the sensitivity and specificity of the prescreen arm are given as follows:

$$\text{Sens}_{\text{prescreen\_arm}} = \text{Sens}_{\text{AI}} \times \text{Sens}_{\text{readers}} \tag{2}$$

$$\text{Spec}_{\text{prescreen\_arm}} = \text{Spec}_{\text{AI}} + (1-\text{Spec}_{\text{AI}}) \times \text{Spec}_{\text{readers}}$$

In the prescreen use case assuming no shift in reader performance when human readers interpret cases that they know were "unknown" by AI, the prescreen arm's sensitivity cannot be greater than the control

Table 2
[a]Illustration of reading order randomization scheme for testing a prescreen AI algorithm.

Session and Block #	Case ID #	Reading Mode
Pre-Study	1–250	AI algorithm classifies cases as “negative” or “unknown”
Session 1, Block 1	1–125	Standard interpretation (control arm)
Session 1, Block 2	The subset of cases from 126 to 250 classified as “unknown”	Standard interpretation but with knowledge that cases were classified as “unknown” (prescreen arm)
Wash-out Period		
Session 2, Block 1	126–250	Standard interpretation (control arm)
Session 2, Block 2	The subset of cases from 1 to 125 classified as “unknown”	Standard interpretation but with knowledge that cases were classified as “unknown” (prescreen arm)

[a] Ideally, different readers would be assigned different random subgroups of cases in different randomized order, varying the start with prescreen cases then control cases.

arm's sensitivity, while there is a potential for improvement in specificity. Metrics that take both sensitivity and specificity into account are advantageous because the net result of any loss in sensitivity and gain in specificity can be assessed. Some possible summary metrics applicable to the lung cancer study are (i) negative predictive value (NPV), (ii) negative likelihood ratio (NLR), and (iii) the area under the ROC curve (AUC). The NLR and NPV of the prescreen arm are defined in Equations (3) and (4).

$$\mathrm{NLR}_{\mathrm{prescreen\_arm}} = \frac{\{1 - \mathrm{Sens}_{\mathrm{prescreen\_arm}}\}}{\mathrm{Spec}_{\mathrm{prescreen\_arm}}} = \frac{\{1 - (\mathrm{Sens}_{\mathrm{AI}} \times \mathrm{Sens}_{\mathrm{readers}})\}}{\{\mathrm{Spec}_{\mathrm{AI}} + (1 - \mathrm{Spec}_{\mathrm{AI}})(\mathrm{Spec}_{\mathrm{readers}})\}} \tag{3}$$

$$\mathrm{NPV}_{\mathrm{prescreen\_arm}} = \frac{(1 - \mathrm{Prev})(\mathrm{Spec}_{\mathrm{AI}} + [1 - \mathrm{Spec}_{\mathrm{AI}}]\mathrm{Spec}_{\mathrm{readers}})}{(1{-}\mathrm{Prev})(\mathrm{Spec}_{\mathrm{AI}} + [1 - \mathrm{Spec}_{\mathrm{AI}}]\mathrm{Spec}_{\mathrm{readers}}) + \mathrm{Prev}(1 - \mathrm{Sens}_{\mathrm{AI}}\mathrm{Sens}_{\mathrm{readers}})} \tag{4}$$

It's important that the performance metric chosen for the study be sensitive to changes in the AI's standalone accuracy. For example, as the AI's sensitivity and/or specificity increase, the accuracy in the prescreen arm should increase. Also, the performance metric chosen should not be strongly affected by shifts in readers' positivity threshold when interpreting a worklist of “unknown” cases because these shifts may be reader-dependent and evolve over time.

Fig. 3 illustrates the effect of the standalone accuracy of AI on the difference in accuracy between the prescreen and control arms for the lung CT prescreen example, assuming no shift in the readers' positivity threshold. The sensitivity and specificity of the AI algorithm are set at 0.95–0.99 and 0.1–0.5, respectively, based on preliminary validation studies. The prescreen arm's AUC is superior to the control arm's AUC except when AI's sensitivity is 0.95 and specificity is < 0.3. This is not true, however, for NPV and NLR. In particular, the NPV is near zero unless AI has very high sensitivity, where it shows only a marginal improvement in the prescreen arm. Note that this is in part due to the low prevalence in screening situations (set at 4% our example). With such a low prevalence, the NPV in both the prescreen and control arms is quite high and there is very little room for improvement in the prescreen arm (even with large gains in specificity). Because the NPV is affected negligibly by improvements in the AI's standalone accuracy, it may not be a good choice for testing this prescreen AI algorithm. The NLR, which decreases as accuracy improves, is superior in the prescreen arm only when AI has very high sensitivity. This metric is sensitive to improvements in the AI's standalone accuracy (i.e. the difference becomes smaller as the AI improves), but the AI would need better standalone accuracy if the NLR is the chosen performance metric instead of the AUC in order to show the benefit of the prescreen AI algorithm.

Now we consider the situation where readers' performance is affected when interpreting a worklist of “unknown” cases. We assume that readers perform on the same ROC curve with and without the AI (since the AI algorithm is not helping them identify diseased subjects) and that the new worklist shifts the readers' threshold for calling cases negative and positive to the right (i.e. higher sensitivity/lower specificity). Fig. 4 illustrates the behavior of various accuracy endpoints as a function of the accuracy of lung CT prescreen AI in this situation. The NPV again is only slightly affected by the AI standalone performance, thus is very insensitive. Because we're assuming the impact of the prescreen AI is simply to shift the readers' decision threshold (and the ROC curve provides a summary across thresholds), there is no impact to the ROC curve. The AUC, therefore, behaves consistently in Figs. 3 and 4. The NLR, however, which is a measure of accuracy at a particular threshold, is affected dramatically by the shift in readers' positivity threshold. Thus, we conclude that the AUC is a good choice as an accuracy metric for testing the lung CT prescreen AI algorithm because it is sensitive to improvements in the AI's accuracy, yet is not affected by changes in shifts in readers' positivity thresholds.

4.1. Comparison with CAD studies

CAD algorithms often increase sensitivity at a cost of specificity, so a net measure of accuracy such as the AUC is important for CAD studies as well as prescreen studies. CAD studies often employ a probability or ordinal rating scale to capture readers’ confidence in the presence of disease. Because the goal of CAD studies is often to assess whether the algorithm can be used to improve reader accuracy at the lesion-level (i.e. improve their ability to identify cancerous lesions and dismiss non-cancerous lesions), the confidence score is assigned to each suspicious lesion identified by the reader. In contrast, for the prescreen AI algorithm this is not necessary since the algorithm is not helping readers find lesions. Thus, for the prescreen AI algorithm study, readers can simply assign a confidence score to the case, rather than at the lesion level.

5. Sample size considerations

For testing a prescreen AI algorithm for lung CT images, the sensitivity and specificity of human readers without the AI algorithm (control arm) is expected to be 0.938 and 0.734, respectively [4]. The AUC of a ROC curve with a FPR = 1-0.734 and Sens = 0.938, and assuming a binormal model with binormal parameters A = 2.16 and B = 1, is 0.937 [8]. A test of non-inferiority is planned, where the null hypothesis is that readers' mean AUC with the prescreen AI algorithm is less than their mean AUC without the prescreen AI algorithm; the alternative hypothesis is that readers’ mean AUC with the prescreen AI algorithm is not inferior. A non-inferiority margin of 0.05 will be used.

The minimum performance of the prescreen AI algorithm in order to achieve equivalence in the magnitude of the AUC between the prescreen and control arm can be determined. We refer to this as the *equivalence boundary*, which provides a useful target in developing the AI algorithm, as well as planning sample size. Fig. 5 illustrates the equivalence boundary for this study for three possible shapes of the ROC curve. If the standalone specificity of the AI algorithm is 0.50 and parameter B = 1.0, then its sensitivity must be ≥ 0.91. If the standalone specificity is only 0.10, then the sensitivity must be ≥ 0.99.

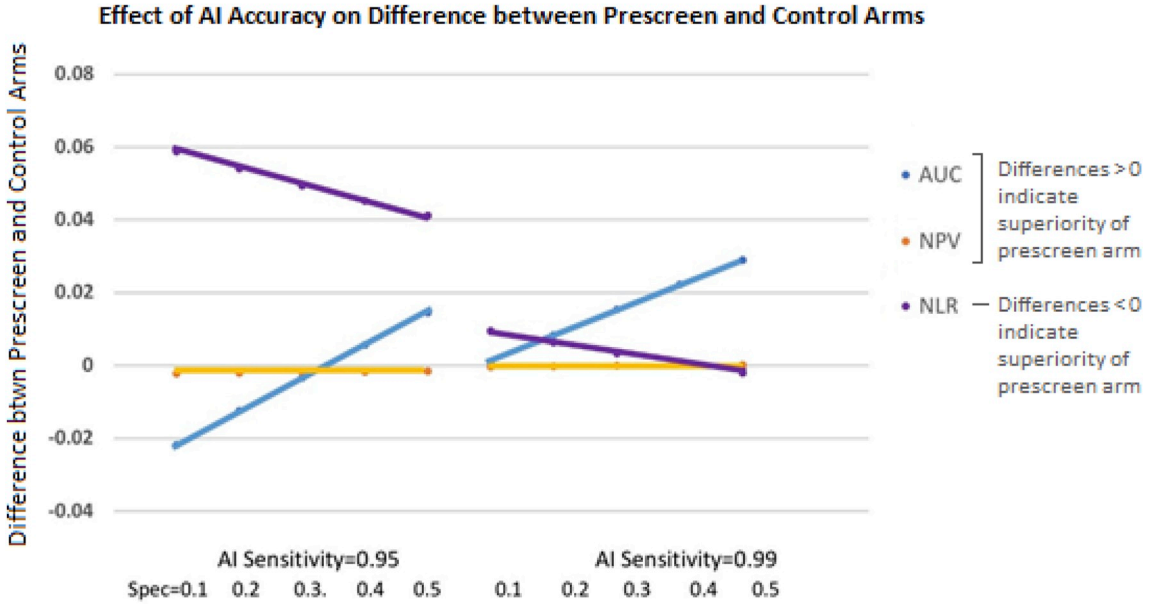

Fig. 3. Difference in accuracy between control and prescreen arms for various accuracy metrics as a function of the standalone performance of the AI device. AI standalone sensitivity is illustrated as 0.95 or 0.99, and its standalone specificity is 0.1, 0.2, 0.3, 0.4, and 0.5. The human reader sensitivity and specificity are set at 0.938 and 0.734, respectively, with disease prevalence of 4%. In the control arm, for the area under the ROC curve (AUC), at a FPR = 1-0.734 and Sens = 0.938, and assuming a binormal model with binormal parameter B = 1, we determined that binormal parameter A = 2.16 (based on Sensitivity = $\Phi(A + B\Phi^{-1}(FPR))$) [8]. Other parameterizations of the ROC curve will give different results. For AUC and NPV, a positive-valued difference (as illustrated on the y-axis) suggests higher accuracy in the prescreen arm than the control arm; for the NLR a negative-valued difference suggests improved accuracy in the prescreen arm.

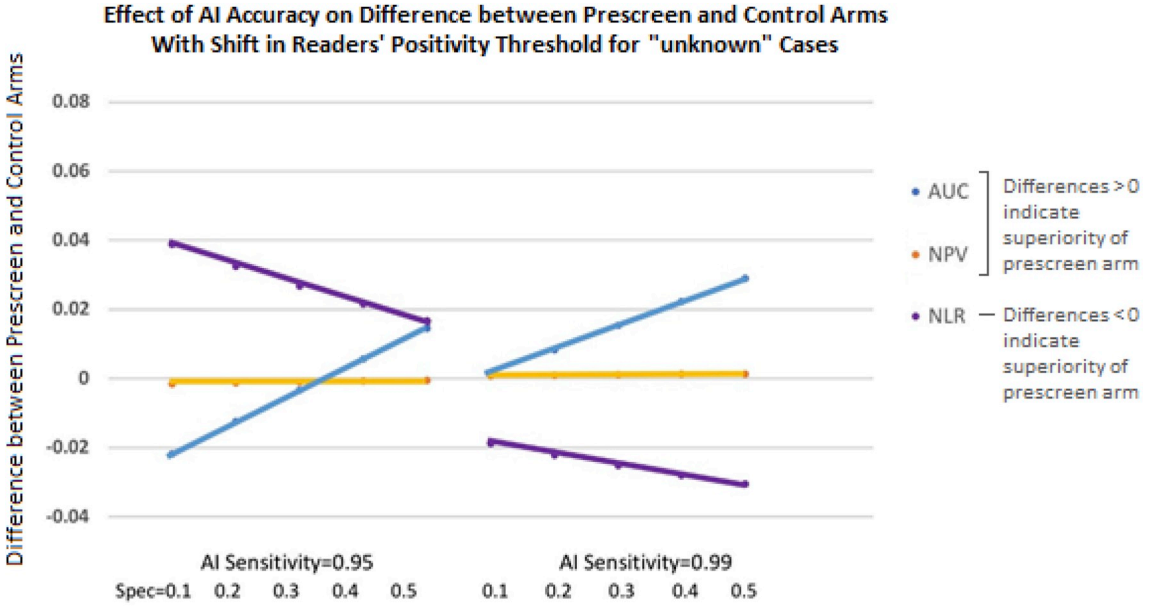

Fig. 4. Difference in accuracy between control and prescreen arms for various accuracy metrics as a function of the standalone performance of the AI device; here readers' shift to a lower threshold for calling cases positive when interpreting AI's "unknown" cases. AI standalone sensitivity is either 0.95 or 0.99, and its standalone specificity is 0.1, 0.2, 0.3, 0.4, and 0.5. For the AUC, we supposed that when readers are interpreting cases classified as "unknown" by the AI algorithm, they shift their specificity from 0.734 to 0.634; if maintaining the same ROC curve (i.e. binormal model with A = 2.16 and B = 1), the corresponding sensitivity is 0.966. For AUC and NLR, a positive-valued difference (as illustrated on the y-axis) suggests higher accuracy in the prescreen arm than the control arm. Note that different magnitudes of readers' shift in threshold for calling cases positive (less shift or more shift) will change the metrics accordingly (less change or more change, respectively).

Developers of the prescreen AI algorithm for lung CT images estimate its standalone sensitivity and specificity to be 0.94 and 0.30 (indicated with an asterisk in Fig. 5). From equation (2), the sensitivity and specificity in the prescreen arm are expected to be 0.88 and 0.81. The AUC of a ROC curve with a FPR = 1-0.81 and Sens = 0.88, and assuming a binormal model with binormal parameters A = 2.05 and B = 1, is 0.929. Sample size methods have been developed that take into account both the number of cases needed (e.g. number with and without lung cancer) and number of readers needed in order to detect a specified difference between two modalities or to test a specific non-inferiority hypothesis [8,9]. There is a trade-off between the number of readers and number of cases needed, such that investigators can choose a design with more readers and fewer cases or few readers and more cases, depending on available resources. For our lung CT prescreen AI

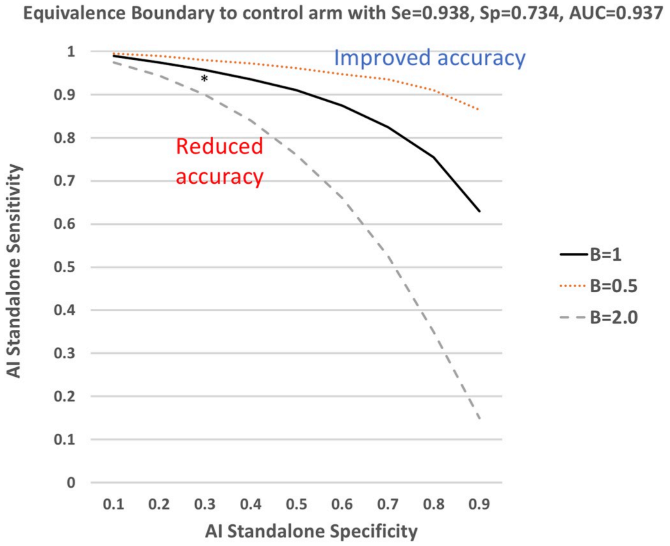

Fig. 5. Equivalence boundary expressed as the prescreen AI's standalone specificity (x-axis) and sensitivity (y-axis) corresponding to a control arm with AUC = 0.937. Three boundaries are displayed for three ROC curve shapes: B = 0.5, 1.0, and 2.0. Above the boundary, the performance in the prescreen arm is superior to the performance in the control arm, while below the boundary the performance is inferior in the prescreen arm. The asterisk indicates the AI developers' estimate of the accuracy of the algorithm. The equivalence boundary was calculated assuming a binormal ROC curve model [8] and assuming parameter B is the same in the control and prescreen groups. For pairs of AI standalone sensitivities and specificities, the sensitivity and specificity in the prescreen arm were calculated from Equation (2), then parameter A and the AUC were estimated from $A = \Phi^{-1}(sens) - \mathrm{B}\times\ \Phi^{-1}(1 - spec)$ and $AUC = \Phi(A/\sqrt{(1 + B^2)})$, where Φ is the cumulative distribution function of a standard normal random variable and Φ^{-1} is its inverse.

algorithm example, we might assume that readers perform on the same ROC curve as in the control arm when interpreting the “unknown” cases (i.e. we do not need to be concerned about shifts in readers’ thresholds). Then, based on a difference in AUCs in the prescreen and control arms of 0.929–0.937 = −0.008, and using a non-inferiority margin of 0.05, one potential sample size combination for 80% power and 5% type I error rate is 10 readers, 125 subjects with lung cancer, and 125 subjects without lung cancer. Ideally, one would perform a pilot study to investigate the assumptions underlying the sample size calculations prior to planning a large pivotal trial.

5.1. Comparison with CAD studies

In studies testing CAD algorithms, it is not possible to directly translate the standalone performance of the CAD into the expected performance of readers in the prescreen arm because it's not known whether readers will accept CAD TP marks or how often they will dismiss CAD FP marks. Instead, investigators often perform a pilot study to estimate the improvement in readers' performance with CAD or take an educated guess. In contrast, for a prescreen AI algorithm the performance of readers in the prescreen arm can be estimated directly from the standalone performance of the algorithm. Thus, determining sample size for testing a prescreen AI algorithm requires fewer assumptions than determining sample size for a CAD study.

6. Discussion

The literature is rich with methodology and applications of CAD as first- and second-readers [8,10,11]. This methodology has been used to assess AI first- and second-reader algorithms. For example, Hale et al. [12] assessed the performance of machine learning models to differentiate meningioma grades on MRI images using standard diagnostic accuracy performance metrics, and similarly Rodriguez-Ruiz et al. [13] used standard measures of accuracy to compare readers’ performance with and without AI for detecting breast cancer. Prescreening AI algorithms are a relatively new application. This application has the potential to greatly reduce the time burden to radiologists and to direct them to more difficult cases. We found that many of the standard CAD study design strategies and performance metrics are applicable, with minor modifications, to testing prescreening AI algorithms.

Several authors have provided general guidelines for assessing AI algorithms used in medical image interpretation. Parmar et al. [14] describe “best practices” and pitfalls for designing and analyzing imaging studies with AI. Parmar et al. [14], Park and Han [2], and Kim et al. [3] discuss the need for adequate sample sizes for both training and validation, locked independent external testing cohorts collected prospectively from preferably multiple sites to avoid biases, proper blinding of testing cohorts during training as well as while tuning of the algorithm, and reporting multiple performance metrics (sensitivity, specificity, AUC, NPV). Park and Han [2] emphasize the need for a spectrum of disease manifestations for evaluating the AI with external datasets, including temporal and geographic variations, to avoid overfitting and overparameterized algorithms. They suggest assessment of both discrimination and calibration performance, when applicable. These recommendations seem equally applicable to AI algorithms used in first- and second-reader, triage, and prescreen modes.

Some authors have used agreement studies to assess the validity of AI algorithms. The study by Bien et al. [15] is such an example where the AI findings on knee MRIs were compared against the majority decision of experts on both an internal and external validation dataset. Savadjiev et al. [16] point out that comparing AI's findings to experts' findings can lead to results that are unreliable because of the inherent subjectivity of the interpretations. They encourage investigators to use patient outcomes and reference standards for comparisons with the AI findings. For testing prescreen AI algorithms, we also noted strong limitations with agreement studies and thus agree with the recommendations of Savadjiev et al.

With respect to measures of diagnostic accuracy performance, the AUC is commonly used in first- and second-reader CAD studies, and we found it applicable for testing prescreen AI algorithms as part of a prospective reader study. For retrospective studies where only binary decisions are available, NPV and NLR are possible options. Biggerstaff [17] has shown that superiority of one modality over another in negative likelihood ratio (NLR) implies superiority in negative predictive value (NPV). With the dependence of the NPV on prevalence, and given its poor responsiveness to improvements in the standalone performance of the AI algorithm, we found that the NLR was a better choice than NPV. However, we also showed that the NLR is strongly affected by shifts in readers' cutpoints for calling cases positive vs. negative, as might occur when interpreting “unknown” images after prescreening. Although calculation of the AUC requires ordinal confidence scores, it allows investigators to provide a summary measure across thresholds and therefore offers protection against the aforementioned shifts. Until there is more research to examine these and other possible metrics in a variety of situations (i.e. broader range of human readers' and AI's sensitivity and specificity, broader range of disease prevalence, range of magnitudes for the inter-reader variability, and effect of the binormal model for the ROC curve) and to understand the effect of limiting radiologists' worklists to cases classified by a prescreen AI algorithm as “unknown”, the AUC from a prospective study seems to be a sound choice.

7. Conclusion

In designing a study to test the diagnostic accuracy of a new prescreen AI algorithm for lung cancer screening images, we first considered a simple retrospective agreement study, and, after evaluating strengths and weaknesses, determined that a multi-reader ROC study was needed. This process allowed investigators to understand why the more complicated study design is warranted. We identified future research areas, including the need to evaluate if and how reader behavior is impacted by knowledge of prescreen AI results. Further methodologic research is needed to address study design and analysis of the various applications of AI algorithms, particularly for assessing efficacy at the patient outcome and societal levels.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke Vasc Neurol 2 (2017) 230–243.

[2] S.H.P. Park, K. Han, Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction, Radiology 286 (2018) 800–809.

[3] D.W. Kim, H.Y. Jang, K.W. Kim, Y. Shim, S.H. Park, Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers, KJR 20 (2019) 405–410.

[4] The National Lung Screening Trial Research Team, Results of initial low-dose computed tomographic screening for lung cancer, N. Engl. J. Med. 368 (2013) 1980–1991.

[5] M.S. Pepe, T.A. Alonzo, Comparing disease screening tests when true disease status is ascertained only for screen positives, Biostatistics 2 (2001) 249–260.

[6] S. Schalekamp, B. van Ginneken, E. Koedam, M.M. Snoeren, A.M. Tiehuis, R. Wittenberg, N. Karssemeijer, C.M. Schaefer-Prokop, Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images, Radiology 272 (2014) 252–261.

[7] M. Meziane, P. Mazzone, E. Novak, M.L. Lieber, O. Lababede, M. Phillips, N.A. Obuchowski, A comparison of four versions of a computer-aided detection system for pulmonary nodules on chest radiographs, J. Thorac. Imaging 27 (2012) 58–64.

[8] X.H. Zhou, N.A. Obuchowski, D.L. McClish, Statistical Methods in Diagnostic Medicine, second ed., Wiley and Sons, Inc., New York, 2011.

[9] W. Chen, N.A. Petrick, B. Sahiner, Hypothesis testing in noninferiority and equivalence MRMC ROC Studies, Acad. Radiol. 19 (2012) 1158–1165.

[10] B.D. Gallas, H.P. Chan, C.J. D'Orsi, L.E. Dodd, M.L. Giger, D. Gur, E.A. Krupinski, C.E. Metz, K.J. Myers, N.A. Obuchowski, B. Sahiner, A.Y. Toledano, M.L. Zuley, Evaluating imaging and computer-aided detection and diagnosis devices at the FDA, Acad. Radiol. 19 (2012) 463–477.

[11] L.H. Eadie, P. Taylor, A.P. Gibson, A systematic review of computer aided diagnosis in diagnostic cancer imaging, Eur. J. Radiol. 81 (2012) 70–76.

[12] A.T. Hale, D.P. Stonko, L. Wang, M.K. Strother, L.B. Chambless, Machine learning analyses can differentiate meningioma grade by features on magnetic resonance imaging, Neurosurg. Focus 45 (2018) 1–6.

[13] A. Rodriguez-Ruiz, E. Krupinski, J.J. Mordang, K. Schilling, S.H. Heywang-Kobrunner, I. Sechopoulos, R.M. Mann, Detection of breast cancer with mammography: effect of an artificial intelligence support system, Radiology 00 (2019) 1–10.

[14] C. Parmar, J.D. Barry, A. Hosny, J. Quackenbush, H.J.W.L. Aerts, Data analysis strategies in medical imaging, Clin. Cancer Res. 24 (2018) 3492–3499.

[15] N. Bien, P. Rajpurkar, R.L. Ball, et al., Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet, PLOS Nov 27 (2018).

[16] P. Savadjiev, J. Chong, A. Dohan, M. Vakalopoulou, C. Reinhold, N. Paragios, B. Gallix, Demystification of AI-driven medical image interpretation: past, present and future, Eur. Radiol. (Aug 2018) published online.

[17] B.J. Biggerstaff, Comparing diagnostic tests: a simple graphic using likelihood ratios, Stat. Med. 19 (2000) 649–663.