Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians (CoDoC)

Krishnamurthy Dvijotham (✉ dvij@cs.washington.edu)
Google

Jim Winkens
Google

Melih Barsbey
DeepMind

Sumedh Ghaisas
DeepMind

Nick Pawlowski
Microsoft

Robert Stanforth
DeepMind

Patricia MacWilliams
Google

Zahra Ahmed
DeepMind

Shekoofeh Azizi
Google https://orcid.org/0000-0002-7447-6031

Yoram Bachrach
DeepMind

Laura Culp
Google

Mayank Daswani
Google

Jan Freyberg
Google

Christopher Kelly
Google Health https://orcid.org/0000-0002-1246-844X

Atilla Kiraly
Google (United States) https://orcid.org/0000-0002-6613-3581

Scott McKinney

OpenAI

Basil Mustafa

Google

Vivek Natarajan

Google

Krzysztof Geras

NYU Grossman School of Medicine https://orcid.org/0000-0003-0549-1446

Jan Witowski

NYU Grossman School of Medicine https://orcid.org/0000-0001-9284-4830

Zhi Zhen Qin

StopTB

Jacob Creswell

StopTB

Shravya Shetty

Google Health https://orcid.org/0000-0003-3783-3172

Marcin Sieniek

Google Inc.

Terry Spitz

Google Health, London https://orcid.org/0000-0002-9791-3767

Greg Corrado

Google, Inc

Pushmeet Kohli

DeepMind

Taylan Cemgil

DeepMind

Alan Karthikesalingam

Google

Article

Keywords:

Posted Date: November 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2231672/v1

Abstract

Diagnostic AI systems trained using deep learning have been shown to achieve expert-level identification of diseases in multiple medical imaging settings[1,2]. However, such systems are not always reliable and can fail in cases diagnosed accurately by clinicians and vice versa[3]. Mechanisms for leveraging this complementarity by learning to select optimally between discordant decisions of AIs and clinicians have remained largely unexplored in healthcare[4], yet have the potential to achieve levels of performance that exceed that possible from either AI or clinician alone[4].

We develop a Complementarity-driven Deferral-to-Clinical Workflow (CoDoC) system that can learn to decide when to rely on a diagnostic AI model and when to defer to a clinician or their workflow. We show that our system is compatible with diagnostic AI models from multiple manufacturers, obtaining enhanced accuracy (sensitivity and/or specificity) relative to clinician-only or AI-only baselines in clinical workflows that screen for breast cancer or tuberculosis. For breast cancer, we demonstrate the first system that exceeds the accuracy of double-reading with arbitration (the “gold standard” of care) in a large representative UK screening program, with 25% reduction in false positives despite equivalent true-positive detection, while achieving a 66% reduction in clinical workload. In two separate US datasets, CoDoC exceeds the accuracy of single-reading by board certified radiologists and two different standalone state-of-the-art AI systems, with generalisation of this finding in different diagnostic AI manufacturers. For TB screening with chest X-rays, CoDoC improved specificity (while maintaining sensitivity) compared to standalone AI or clinicians for 3 of 5 commercially available diagnostic AI systems (5–15% reduction in false positives). Further, we show the limits of confidence score based deferral systems for medical AI, by demonstrating that no deferral strategy could have achieved significant improvement on the remaining two diagnostic AI systems.

Our comprehensive assessment demonstrates that the superiority of CoDoC is sustained in multiple realistic stress tests for generalisation of medical AI tools along four axes: variation in the medical imaging modality; variation in clinical settings and human experts; different clinical deferral pathways within a given modality; and different AI softwares. Further, given the simplicity of CoDoC we believe that practitioners can easily adapt it and we provide an open-source implementation to encourage widespread further research and application.

1 Introduction

Deep learning-based AI systems achieve impressive accuracy in many applications, but the lack of a "safety net" impacts deployment in safety-critical areas where the consequences of AI error often mean that staged use alongside human experts is crucial. One such safety-critical application is medical imaging, where diagnostic AI systems have demonstrated expert performance in multiple retrospective research studies[1], but where AI models can make errors in cases that can be diagnosed accurately by clinicians.

Although there is a rich literature of AI (and other computerised technologies for medical imaging diagnosis) providing diagnostic outputs to assist clinicians and potentially improve their performance [5–8], it remains unclear how to optimally resolve situations in which such AI tools and human experts have similar levels of performance, but have opposing diagnostic opinions where one must be chosen. Patient harm could theoretically arise from either the choice to override a clinician's independent opinion with AI or vice-versa[4]. Optimal care requires deference to the diagnostic opinion that is most likely correct, but this is challenging to predict because failure modes of AI systems have been difficult to characterise. Resolving this situation with a reliable method for optimally choosing which diagnostic agent to defer to therefore represents an important unmet need for the application of AI to healthcare.

As diagnostic AI tools and clinicians have been shown to make errors in different types of cases[9,10], an ideal predictive system might harness their complementary strengths. We present a Complementarity driven Deferral-to-Clinical Workflow (CoDoC) system to improve performance of AI-only and human-only clinical workflows, by first examining the predictions of a diagnostic AI system and then deferring to a clinical workflow if the AI is deemed to be less likely to be accurate than the clinical workflow for that case. As complementarity between AI and clinicians has been observed in multiple medical settings, we enable re-implementation of our framework in other settings by open-sourcing our model code alongside the clinical data required to reproduce our experimental results.

We leverage algorithmic approaches enabling AI tools to defer to domain experts when uncertain about their predictions and show that this can increase the performance of the composite system by optimising reliance on the correct inference. However, most precedent work in the ML literature on learning to defer to a human collaborator is not applicable in medical AI where regulatory requirements, engineering, data-sharing or intellectual property considerations may require the diagnostic AI to be accessible only as a "locked model" that cannot be modified [11, 12,13] (a black-box setting). Given this practical constraint, we develop a system that is compatible with any pre-existing diagnostic AI model without requiring it to be retrained. Our system uses confidence scores from one or more "locked" (pretrained) diagnostic AI models as inputs to a "deferral AI" model that decides whether to make a prediction using the diagnostic AI models or defer to a clinician; and can thereby be implemented as a wrapper.

From an engineering perspective, this enables a modular system design and ensures that the deferral AI can be simple as it only learns which confidence scores of the prediction model should result in a deferral to a clinician. As a consequence, the CoDoC system is nimble and can be adapted to novel deployment scenarios and any diagnostic AI. The deferral AI is trained based on simple non-iterative algorithms. Despite the simplicity of our approach, we show that CoDoC is effective and given a small tuning set of cases to train the deferral AI, can attain performance that exceeds that of diagnostic AI or clinicians alone across multiple diagnostic tasks and clinical workflows, while being robust to various forms of distribution shift. Further improvements may be possible by co-training the diagnostic and deferral AI, as has been suggested in prior work[14]. However, restricting ourselves to treating the diagnostic AI as a black box also has several advantages, including the ability to use third-party diagnostic models and avoid contravening regulatory/IP restrictions on modifying the diagnostic AI.

2 Description Of The Codoc System

Once trained, a deployed CoDoC system takes as its input the confidence score from the diagnostic AI system for a given medical image. This score is fed into a deferral AI model which decides to either use the diagnostic model or defer, in which case the medical image is diagnosed using a standard clinical workflow without diagnostic AI. Details of the training and deployment architecture of CoDoC are described in Fig. 1.

The learnable component of CoDoC is the deferral AI. The deferral AI is inspired by work showing that the confidence score from a deep network[15] can be used to detect inputs where an AI model's predictions are unreliable. This was also demonstrated to be effective in medical AI systems for breast cancer screening in recent work[16]. In these prior works, the confidence score is taken as an indication of whether the prediction is reliable, and these works compute thresholds on the confidence score under which to defer to a clinician, so as to maximally boost the performance of the overall system. We refer to this approach as *threshold search*. However, in most practical applications, the amount of data available for optimising the thresholds is very small (particularly in terms of the number of positive cases) and the estimates of combined performance obtained can be very noisy and fail to generalise to future unseen data.

In particular, consider the plot from Fig. 1d where, for a given operating point that a diagnostic AI model operates at, we plot points (synthetically generated for illustration purposes) with markers identifying where the clinician predicts the correct ground truth label and where the model predicts the correct ground truth label. A threshold search approach faces an underspecification problem[17] as shown by the first three panels of Fig. 1d - there are many different choices for the thresholds on the confidence score of the diagnostic AI within which to defer to a clinician that all achieve the same performance (as they all lead to deferring on all cases where the diagnostic AI made an incorrect prediction while the clinician made a correct prediction, and not deferring on cases where the diagnostic AI is correct but the clinician is incorrect). How to choose between these solutions remains an open question. In the next section, we derive a mathematically optimal deferral strategy and an approach that addresses the underspecification problem.

2.1 Mathematically Optimal Deferral strategy and Comparison with Threshold Search

We used a probabilistic approach to estimate the performance of the deferral AI given a limited dataset that was separate from the dataset the diagnostic AI was trained on. We call this the “tuning” dataset and it was the dataset that the deferral AI was trained on. Each datapoint in the tuning dataset comprised a three-tuple of the model confidence score, the ground truth disease label (result of a biopsy for Cancer or an Xpert MTB/RIF assay for TB[18]) and the result of a clinical workflow assessing the case. We provide details of the same in Appendix Section A.1 and a high level summary of the approach here. Given an operating point θ for the diagnostic AI (a threshold where for confidence outputs z are classified as AIOpinion = True), the specificity and sensitivity of any deferral strategy can be estimated by computing an integral based on this density estimate, as follows:

$$\begin{aligned}
\text{Sensitivity} &= P(\text{CoDoCPrediction} = \text{True}|\text{GroundTruth} = \text{True}) \\
&= \int_0^1 P(\text{AIScore} = z, \text{ClinicianOpinion} = \text{True}|\text{GroundTruth} = \text{True})\text{Defer}(z)dz \\
&\quad + \int_0^1 P(\text{AIScore} = z, \text{AIOpinion} = \text{True}|\text{GroundTruth} = \text{True})(1 - \text{Defer}(z))dz \\
\text{Specificity} &= P(\text{CoDoCPrediction} = \text{False}|\text{Ground Truth} = \text{False}) \\
&= \int_0^1 P(\text{AIScore} = z, \text{ClinicianOpinion} = \text{False}|\text{GroundTruth} = \text{False})\text{Defer}(z)dz \\
&\quad + \int_0^1 P(\text{AIScore} = z, \text{AIOpinion} = \text{False}|\text{GroundTruth} = \text{False})(1 - \text{Defer}(z))dz
\end{aligned}$$

where

$$\begin{aligned}
&\text{Defer}(z) = 1 \text{ indicates CoDoC defers when AIScore} = z \\
&\text{Defer}(z) = 0 \text{ indicates CoDoC does not defer when AIScore} = z
\end{aligned}$$

Moreover, the AIOpinion is simply based on the operating pointθ

$$\begin{aligned}
P(z, \text{AIOpinion}|\text{GroundTruth}) &= P(\text{AIOpinion}|z) \cdot P(z|\text{GroundTruth}) \\
P(\text{AIOpinion} = \text{True}|z) &= \text{IsLess}(\theta, z) \\
P(\text{AIOpinion} = \text{False}|z) &= (1 - \text{IsLess}(\theta, z)) = \text{IsLess}(z, \theta)
\end{aligned}$$

where IsLess(a, b) is a function that is one when a < b and zero otherwise. Now, we choose a parameter λ between 0 and 1 that indicates how to trade-off sensitivity and specificity, and we train CoDoC to optimise

$$\lambda \cdot \text{Sensitivity} + (1 - \lambda) \cdot \text{Specificity}$$

Based on this objective, in the appendix we show that we can define the following advantage function for any value of z as

$$\begin{aligned}
\text{Advantage}(z) = \lambda &\cdot P(\text{AIScore} = z, \text{ClinicianOpinion} = \text{True}|\text{GroundTruth} = \text{True}) \\
&+ (1 - \lambda) \cdot P(\text{AIScore} = z, \text{ClinicianOpinion} = \text{False}|\text{GroundTruth} = \text{False}) \\
&- \lambda \cdot P(\text{AIScore} = z|\text{GroundTruth} = \text{True}) \cdot \text{IsLess}(\theta, z) \\
&- (1 - \lambda) \cdot P(\text{AIScore} = z|\text{GroundTruth} = \text{False}) \cdot \text{IsLess}(z, \theta)
\end{aligned}$$

For any choice of operating point θ for the AI, the optimal deferral strategy is:

$$\text{Defer}(z) = \begin{cases} 1 & \text{if Advantage}(z) > 0 \\ 0 & \text{otherwise} \end{cases}$$

We show mathematically that this is the optimal deferral strategy in Appendix A.1.1.1.

Since the AI prediction is obtained by checking whether the AI Score is above the operating point, the Advantage function is completely determined by the 4 conditional distributions:

$$P(\text{AIScore}, \text{ClinicianOpinion} = \text{True/False} | \text{GroundTruth} = \text{True/False})$$

We can estimate the above conditional distributions using any density estimation method like Kernel Density Estimation[19]. Using a density estimation method leads to a smoothing effect on the advantage function that is critical for generalisation in data sparse settings. In comparison, threshold search can be interpreted as applying the same strategy without any smoothing on the conditional distributions, simply taking these to be the empirical distributions based on the dataset. In Fig. 1d, we plot the comparison of the advantage functions from the threshold search approach and show that it leads to an underspecification problem, i.e., there are many choices for the thresholds that all achieve the same performance in terms of sensitivity and specificity on the tuning dataset. However, when we use the smoothing approach and compute the advantage function, the smoothed advantage function has unique zero crossings that define the deferral region.

The density estimation method we used naturally has hyperparameters that control the level of smoothing. The ideal amount of smoothing is dataset-dependent and cannot be chosen a-priori. We instead used a hold out validation set to choose hyperparameters of the smoothing procedure so that we choose values that compute a deferral strategy that generalises well to unseen data (details in Appendix Section A.1.3).

3 Results Of Codoc On Medical Imaging

We studied two diagnostic tasks for evaluating CoDoC: breast cancer detection from X-ray mammograms and recommending patients for diagnostic sputum testing for TB based on TB-suspicious findings in their Chest X-Ray. We considered multiple international clinical settings: TB screening in Bangladesh, US mammography (where clinicians predominantly operate alone as a “single reader”); and UK mammography where breast screening requires two readers and an arbitration process for each case (“double reading”). In each setting, we demonstrate that CoDoC achieves superior diagnostic performance compared to either diagnostic AI models or clinicians alone (Table 1). The details of each clinical task are given in the subsequent sections.

For each clinical task, the set of all available data was partitioned into a training set and a held-out test set, that was only touched once for evaluating the final CoDoC system. The training set was further partitioned into a tuning set for the diagnostic AI and a validation set that was used to select hyperparameters or select amongst several deferral AI models that performed well on the tuning set. For each setting, we optimised CoDoC either for improving sensitivity keeping specificity intact or vice versa, relative to both the clinical workflow and the diagnostic AI model. The operating point for the diagnostic AI model was chosen to reflect the tradeoff made between sensitivity and specificity by the clinical workflow as follows: for the purposes of comparing reader, diagnostic AI and CoDoC performance, we obtained an operating point that matched the clinical workflow on specificity, another that matched the clinical workflow on sensitivity, and picked the mean of those two as the operating point for the diagnostic AI we compared to, following published precedent[20].

Table 1

Performance of CoDoC compared to a standalone diagnostic AI system and clinical readers. Bolded and starred entries mean statistically significant improvement over both the AI-only and clinician-only baselines (at a statistical significance threshold of .05). Unbolded entries mean statistically non-inferior (with a non-inferiority margin of .05) to both AI-only and clinician only baselines). OP 1–4 refers to a set of distinct parameters that determine an operating point for each application. There were no statistically inferior entries to either AI-only or clinician-only. Superiority and non-inferiority were defined by standardised hypothesis tests (McNemar and Wald; see Appendix section A.1.3 for details).

Task / Dataset	**Diagnostic AI model**	**Clinical Workflow**	**CoDoC**		**Clinician(s)**		**Standalone diagnostic AI model**	
			Sens.	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**
Breast Cancer Detection/ UK Mammo Dataset	Mammo Diagnostic AI 1	Defer to First reader OP 1	**72.6***	91.8	62.7	92.8	64.9	93.9
		Defer to First reader OP 2	64.4	**96.5***				
		Defer to Double Read + Arbitration OP 3	**71.0***	95.1	67.6	96.1		
		Defer to Double Read + Arbitration OP 4	68.3	**96.9***				
Breast Cancer Detection/ US Mammo Dataset 1	Mammo Diagnostic AI 1	Defer to Single Reader OP 1	**56.9***	87.5	50.0	88.0	48.3	86.3
		Defer to Single Reader OP 2	50.0	**91.2***				
Breast Cancer Detection/ US Mamm Dataset 2	Mammo Diagnostic AI 2	Defer to Single Reader	96.7	**90.9**	96.7	84.5	86.7	70.0
TB Detection/ TB Dataset	Manufacturer 1	Defer to Single Reader	90.5	**68.3**	89.4	62.4	90.9	65.6
	Manufacturer 2		89.3	**64.2**			89.5	62.5
	Manufacturer 3		89.2	**67.7**			89.7	62.7

Task / Dataset	Diagnostic AI model	Clinical Workflow	CoDoC		Clinician(s)		Standalone diagnostic AI model	
			Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
	Manufacturer 4		90.6	70.8			91.9	67.9
	Manufacturer 5		91.8	71.7			92.8	69.1

3.1 Breast cancer screening with x-ray mammography

3.1.1 Datasets and Diagnostic AI Systems

We utilised an updated version of a previously published diagnostic AI model ("Mammography Diagnostic AI 1"), trained for breast cancer classification in two large datasets of screening mammograms from the UK ("UK Mammography Development Dataset") and US ("US Mammography Dataset 1")[10], incorporating published technical improvements to maximise AI performance [21,22]. The information in both datasets was de-identified [10].

To enable direct comparison to our previous work we report results on the same held-out test set[10] of UK cases that were not used to train or tune either the diagnostic AI model or the deferral AI model ("UK Mammography Test Dataset"). The data used to train CoDoC was restricted to a subset of the same UK dataset cases used for training the previously published diagnostic AI model (those used for tuning its hyperparameters)[10]. This set consisted of 60755 cases (of which 1198 have biopsy confirmed cancers). This was further split into a training subset consisting of 12169 cases (of which 243 were cancer positive) and a validation subset consisting of the remaining cases. The training subset was used for optimising parameters of the deferral AI and validation subset was used for choosing between a finite collection of deferral AI models trained with different hyperparameters.

The UK held-out test set comprised screening mammograms collected between 2012 and 2015 from 25,856 women at two screening centres in England, where women are screened every three years. This was a random sample of 10% of all women with screening mammograms at these sites during this time period. For each case, the cancer ground truth was determined through biopsy-confirmed longitudinal follow up over 39 months, and historical clinical reads were also collected. Biopsy-confirmed breast cancer was found in 414 women within 39 months of imaging. In the UK, two readers interpret each mammogram and an arbitration process may be used to invoke a third opinion ("double-reading"). These interpretations occur serially, such that each reader has access to the opinions of previous readers. The records of these decisions yield different benchmarks of human performance for cancer prediction, which we included as “First reader” (independent first human reader) and “Double read” (final screening decision, including any arbitration reads as required).

To assess generalisation, we explored the performance of CoDoC in a separate diagnostic AI system (Mammography Diagnostic AI 2), trained for breast cancer classification in a wholly-separate US dataset (US Mammography Dataset 2) by wholly-separate diagnostic AI developers compared to Mammography Diagnostic AI 1[23]. In this setting CoDoC was trained using 7,074 breast exams from the same “US Mammography Dataset 2” cases used for training the previously published “Mammography Diagnostic AI 2”. The held-out test set consisted of 7,074 breast exams of women aged 27 to 92, performed between May and August of 2017 at NYU Langone Health. Cancer was determined by positive biopsy within 120 days of the original study, and according to previously published methodology. For the training and test sets combined, 307 studies contained only benign lesions, 40 only malignant lesions, 21 studies both benign and malignant lesions and the remaining 13,780 exams came from patients who did not undergo a biopsy. Benign and malignant lesions were split evenly across the training and test sets.

3.1.2 Performance of CoDoC in UK Mammography

When using CoDoC autonomously with deferral to a single reader (reader 1), absolute specificity was improved by 2.6% (95% CI 2.3–2.9%; $P = 0.005$ for superiority at non-inferior sensitivity) or absolute sensitivity by 7.7% (95% CI 4.5–10.7%; $P < 0.001$ at non-inferior specificity) compared to a standalone diagnostic AI model, at a cost of 0.15 and 0.34 additional human reads per case respectively.

Improvements were larger still if CoDoC was used autonomously with deferral to a double-reading standard of care. This improved sensitivity compared to double-reading without AI by 3.4% (95% CI 1.3–5.4%; $P < 0.001$ at non-inferior specificity), or specificity by 0.8% (95% CI 0.6-1.0%; $P < 0.001$ for superiority at non-inferior sensitivity) and reduced human reads by 1.4 per case. CoDoC here also significantly improved performance compared to using a diagnostic AI model alone (Table 1).

Clinical interpretation of results

The observed improvements in sensitivity and specificity could have implications for the implementation of AI in screening mammography. We considered a number of potential ways that CoDoC might be implemented into a screening workflow: 1) deferral to a single reader alone, 2) deferral to a double-reader workflow.

Considering the situation where the double-reader system was replaced by CoDoC deferred to a single reader alone, CoDoC attained superior diagnostic accuracy to the historic double-reader system using only 15 percent of a single reader's workload. While retaining the same cancer detection rate as a standalone diagnostic AI system, this reduced the "recall to assessment rate" by 37 percent (from 70 to 44 per 1000 cases).

In an alternative deployment scenario where the desired outcome was to maximise diagnostic accuracy, CoDoC could defer to the standard double-reader workflow. This would deliver a valuable 15% reduction in recall-to-assessment (from 48 to 41 per 1000 cases) while matching the cancer detection rate of double-reading, alongside a three-fold reduction in reader workload (from 2.1 to 0.7 reads per case). This deployment workflow would be in accordance with emerging guidelines for the use of diagnostic mammography AI tools in replacement of the second reader in the UK workflow.

3.1.3 Performance of CoDoC in US Mammography

We evaluated the performance of CoDoC in US Mammography when applied to a wholly separate diagnostic AI system [25] built by a different manufacturer (Mammography Diagnostic AI 2), in a wholly separate dataset from a different clinical context (US Mammography Dataset 2). In this setting CoDoC demonstrated a statistically significant improvement compared to both the existing single-reader system in the US, or a diagnostic AI model alone (Mammography Diagnostic AI 2). CoDoC improved absolute specificity by 6.5% (95% confidence interval (CI) 5.9-7.0%; $P < 0.001$ for superiority at non-inferior sensitivity) despite reducing the requirement for human reads by 53% (from every case being read to every 2.1 cases being read). Compared to the standalone diagnostic AI model, CoDoC showed a statistically significant improvement in absolute specificity of 20.9% (95% CI 19.8–22.0%; $P = 0.019$ for superiority at non-inferior sensitivity), at a cost of 0.47 additional human reads per case.

Clinical interpretation of results

The performance improvements from CoDoC would reduce the recall-to-assessment rate by up to 45 percent (from 159 per 1000 cases to 88 per 1000) while matching the cancer detection rate of the current single-reader system.

3.1.5 Comparison of our Probabilistic Approach and Threshold-Search

We present results comparing our approach with the threshold-search approach proposed in prior work[16]. We found that on datasets where we have a large tuning dataset (like the UK Mammography Dataset or the TB Datasets considered in the next section), the two approaches gave nearly identical results. However, when the tuning dataset was small (the US Mammography dataset 2), we found that the probabilistic approach outperformed the threshold-search approach[16] on unseen test data. Figure 2 shows a comparison between the two approaches and demonstrates that CoDoC outperformed the threshold-search approach when the tuning datasets were small (less than 1000 samples) as in the US Mammography Dataset 2 (NYU).

3.2 Tuberculosis screening with 5 commercially available AI systems

3.2.1 Datasets

CXR images are widely used in a triage process to direct which patients require gold standard investigation for TB by genotyping using the GeneXpert diagnostic assay. The decision to request a Xpert diagnostic assay can be made by a human radiologist, with or without reference to an AI system. Where the triage is performed autonomously by AI systems instead of human radiologists (as supported by [26]), this is usually based on a predefined threshold abnormality score.

We explored the impact of CoDoC when applied to screening for tuberculosis (TB) using chest X-rays as a triaging tool for an Xpert MTB/RIF (Xpert) test. We analysed a previously published study [27] describing individuals aged 15 years or older presenting consecutively to 3 tuberculosis screening centres in Dhaka, Bangladesh between May 2014 and Oct 2016 where they received a digital posterior-anterior chest X-ray (CXR) and an Xpert test. 23,954 individuals were included in analysis after exclusion of those under 15 years of age or without a valid CXR or Xpert test, which were split evenly into CoDoC training and test sets. Note that in this setting, the commercial AI systems are not strictly “Diagnostic AI” tools since the objective here is not to determine the presence of TB from the CXR, but rather to identify sufficient suspicion of TB based on the CXR to justify diagnostic investigation with a Xpert test. For consistency with the other modalities in this manuscript and to differentiate this function from the deferral decision, we term these commercial systems “diagnostic AI” systems.

All chest X-rays were read independently by a group of 3 registered radiologists and 5 commercial diagnostic AI systems, each developed by a different manufacturer. Radiologist reads were dichotomised into possible tuberculosis (highly suggestive of tuberculosis, possibly tuberculosis) or not tuberculosis (abnormal but not tuberculosis, normal). Xpert results were used as the bacteriological evidence and reference standard. The 5 diagnostic AI systems processed anonymised CXR images retrospectively without any clinical or demographic information, independently, and with no previous training or validation at the study site. We assessed the role of CoDoC in a hypothesised workflow common in real-world implementation of AI systems for TB screening, where radiologists and AI tools are both available for the independent interpretation of a CXR image. In our evaluation, CoDoC was used to decide when to use the autonomous AI system and when to invoke a radiologist opinion. The proportion of subsequent Xpert assays saved (with 0% representing the Xpert testing-for-all scenario) reflects a proxy for incremental cost-effectiveness achieved by adding CoDoC.

3.2.2 Performance of CoDoC

Compared to the standalone diagnostic AI models of manufacturers 1, 2 and 3, CoDoC improved absolute specificity by 2.7%, 1.7% and 5.0% (95% CI for 1: 2.2%-3.2%, $P < 0.001$, for 2: 1.2–2.3%, $P < 0.001$ and for 3: 4.5–5.6%, $P < 0.001$ for superiority at non-inferior sensitivity) at a cost of a single human read for every 2.9, 3.7 and 3.8 cases respectively. CoDoC improved the proportion of Xpert tests avoided by 5.2%, 3.3% and 9.5% respectively. For manufacturers 4 and 5, CoDoC did not result in superior performance over the diagnostic AI alone, which was markedly superior to human radiologists without deferral. Despite all 5 commercial AI systems showing significantly better performance than the radiologists, the AI models of manufacturers 1–3 could still be improved with CoDoC.

3.2.4 Clinical interpretation of results

For 3 out of 5 commercially available systems, the ability to route cases between radiologists and AI systems resulted in significant gains. Centres using CoDoC to enhance these AI systems would have increased the proportion of Xpert tests avoided by 6.0% on average. For the two highest-performing AI systems in this evaluation dataset, our work demonstrated that any confidence-based deferral regime (including CoDoC) would not confer any improvement over an AI-alone approach, as was borne out by the results. We present a careful analysis of the limits of confidence based deferral in Section 5 using these results. Centres using CoDoC to augment AI systems would therefore be able to predict when (and when not) to employ a deferral-based strategy (versus using AI alone).

4 Ascertaining The Limits Of Confidence-based Deferral

We further sought to assess whether, even in ideal conditions (with infinite data or no distribution shift), whether any confidence based deferral strategy could have succeeded in improving performance. To do so, we made two changes to the CoDoC system: 1) We allowed multiple deferral regions (as many as ten), allowing the system to learn a very flexible function of the confidence score to determine when to defer. Searching for multiple deferral regions can be done efficiently via a dynamic programming algorithm we describe in Appendix Section A.2) We allowed the system to cheat by training directly on the test set used for evaluation.

Our rationale for this was to remove all possible sources of performance degradation that were due to the limited nature of the deferral AI or the dataset used to train. In this setting, we allowed the deferral AI to learn a very general (almost arbitrary) function of the confidence score, and further allowed the model to be trained directly on the test dataset, removing any possibility of drop in performance due to generalisation issues. The performance of CoDoC in this setting can therefore be seen as a theoretical upper bound on the performance of any deferral strategy that is based only on the confidence score of the diagnostic model, and is a useful tool to assess whether there is any potential complementarity between the diagnostic AI and clinical workflow that can be exploited to improve performance. We see from Fig. 4 that indeed, for the two strongest commercially available TB models we studied, this theoretical performance limit (as plotted in the red dashed line) was very close to the performance of the diagnostic AI, thereby showing that confidence based deferral strategies had very little potential for improving performance in this setting.

5 Generalisation Of Codoc To New Populations And Clinical Readers

The accuracy of diagnostic AI models is vulnerable to "distribution shift" in deployment, meaning a discrepancy between the distribution of data used for training the diagnostic AI and the distribution of data seen in deployment[28]. Common distribution shifts include changes in technology (e.g. variations in imaging equipment, acquisition parameters, and post-processing methods), changes in population and setting (e.g. demographics and disease distribution), and changes in behaviour (e.g. evolving clinical practice and guidelines, differences in sociotechnical context or clinician training between centres or countries). In this section, we examine whether CoDoC generalises well under various shifts, with the goal of stress testing the applicability of CoDoC under real-world deployment scenarios.

5.1 Adaptation to a new population

In the previous sections, we demonstrated how CoDoC improved diagnostic accuracy relative to a standalone AI and clinician(s), when CoDoC was trained using tuning data sampled from the same population of patients as that used to train the underlying diagnostic AI, and where performance was evaluated on held-out samples also from the same population. However, in many settings, the diagnostic AI may not have been trained on the population where it is deployed. In this section, we asked the question: can CoDoC yield accuracy improvements over both a standalone diagnostic AI system and clinicians even on a new population that the diagnostic AI has not been trained on?

We evaluated CoDoC in a scenario where the diagnostic AI model (Mammography Diagnostic AI 1) was presented with cases from a new and previously unseen US population ("US Mammography Dataset 2"). This dataset was obtained from a community screening centre where women are screened every year, and images are read by a single reader who importantly also has access to digital breast tomosynthesis (DBT) when making screening recall decisions. The tuning set used to train the deferral AI consisted of cases from a different set of women from the same population, comprising 8783 cases (of which 139 are cancer positive) collected from 2,926 women drawn from the new population. This was further subdivided into 4392 cases used for training the deferral AI and 4391 cases used for model selection. The held-out test set for evaluation had 8981 cases (of which 174 are cancer positive).

The performance of the standalone AI clearly degraded on this new population as can be seen by comparing the ROC curve for the standalone AI from Fig. 5 with that in Fig. 2. CoDoC demonstrated a statistically significant improvement compared to both the existing single-reader system in the US, or a diagnostic AI model alone. CoDoC improved absolute sensitivity by 6,9% (95% confidence interval (CI) 0.4, 13.2%; $P = 0.033$ for superiority at non-inferior specificity) or improved absolute specificity by 3.2% (95% CI 2.6%, 3.7%; $P = 0.037$ for superiority at non-inferior sensitivity) despite reducing the requirement for human reads by 55% (from 1 read per case to less than 1 read every 2 cases).

The performance improvements from CoDoC using the full tuning data would reduce the recall-to-assessment rate by up to 25% (from 127 per 1000 cases to 96 per 1000) while matching the cancer detection rate of a typical single-reader system. Compared to the standalone diagnostic AI model, CoDoC showed a statistically significant improvement in absolute specificity of 4.9% (95% CI 4.1%, 5.6%; $P = 0.037$ for superiority at non-inferior sensitivity) or an improvement in absolute sensitivity of 8.6% (95% CI 1.9%, 15.2%; $P = 0.011$ at non-inferior specificity), at a cost of 0.55 additional human reads per case.

5.2 Generalizing to previously unseen readers

In real-world deployments, different readers may use CoDoC to those whose opinions were used for training the deferral AI (See Fig. 1 for the training architecture for CoDoC). We measured how the performance of CoDoC generalised to previously unseen readers; and examined the variation in performance of CoDoC for different individuals. To do this, we evaluated the impact of the deferral AI for cases read by each previously unseen reader in the UK Mammography Test Dataset (Figure A0 in Appendix Section A.2). CoDoC improved the sensitivity and/or specificity of every individual reader, with a mean (95% C.I.) improvement of 5.4% sensitivity (2.5%-8.2%) and 0.9% specificity (0.6%-1.1%).

6 Subpopulation Analysis

We analysed how the improvements from CoDoC broke down by subgroups in the population of the UK Mammography Dataset, based on clinically relevant characteristics like age and breast density. In particular, we were interested in a) whether there were systematic differences between the improvements from deferral between subgroups and, b) whether the deferral system was able to leverage such systematic performance disparities between the AI and human readers without explicit knowledge about the individual subgroups where the standalone AI system was significantly worse than readers. Further results on analysis and interpretation of the results is presented in Appendix Section A.3.

We focussed on the UK Mammography dataset 1 and considered the deferral to a single reader, as this was the dataset for which we had the most nuanced subgroup information and the deferral to single reader remains the most likely practical deployment scenario for CoDoC.

Differences in gains from deferral between subgroups: As is seen in Fig. 6, CoDoC consistently improved on the performance of both standalone AI and the human reader across all subgroups of age and breast density in either sensitivity or specificity or both. For some subgroups (like those above age 70), the diagnostic AI was significantly worse than the reader on specificity. In such cases, CoDoC recovered much of the performance loss from the diagnostic AI but was unable to outperform the reader.

CoDoC learned to leverage differences between reader and AI performance: As is seen in Fig. 6, CoDoC deferred more on certain subgroups, for example, those older than 70 years. In doing so, CoDoC automatically identified subgroups where the reader was better than the standalone AI, even though the deferral AI never had access to subgroup information. Further subpopulation analyses are included in the extended materials, with de-identified images of cases deferred by CoDoC accompanied by qualitative post-hoc interpretative notes by a board-certified radiologist.

Discussion

In this study we demonstrate a novel CoDoC system that can learn to decide when to rely on a diagnostic AI system and when to defer to clinical experts or workflows. We evaluated CoDoC in multiple simulated clinical workflows screening for breast cancer or TB and showed that combined AI-clinician performance using CoDoC exceeds that currently possible through either AI or clinicians alone. CoDoC is highly configurable to meet the requirements of specific clinical deployments, and does not require access to the inner workings of the target standalone AI diagnostic model. We believe CoDoC represents a step towards harnessing the complementarity possible between AI and clinical experts, to improve accuracy, trust, and safety in real-world clinical deployments.

It is increasingly becoming apparent that clinicians and AI systems fundamentally assess images differently[10], and that both have different strengths and weaknesses [29,30]. It is therefore intuitive that systems designed to combine aspects of both should lead to improvements in both performance and safety. However, in practice, there is an unmet need to enable users of medical AI systems to know which opinion should prevail when their opinion differs from an AI tool and they are uncertain which should prevail. Furthermore, the ability for an AI system to say "I'm not sure" or "I do not know" is an important capability to ensure safe clinical deployment of this technology[31].

A recent study demonstrated that the paradigm of deferral using threshold-search is a promising approach for managing this unmet need[16]. However, this prior work only explored the solution in one medical condition (breast cancer), one diagnostic AI tool and one clinical workflow (for breast cancer screening in a German dataset). It has hitherto remained unclear whether the promise of deferral might be applicable to multiple medical AI applications, how a deferral algorithm might generalise to diagnostic AI tools from multiple different manufacturers, whether performance would be robust given multiple different clinical workflows, and whether a deferral algorithm could be adaptable to new AI tools or clinical settings with very limited data for site-specific training (as is common in medicine). CoDoC validates the hypothesis that algorithmically-driven deferral between AI and clinical experts might improve composite performance in a wide variety of medical AI applications screening for cancers and TB alike, with rigorous evaluation in multiple countries for multiple different AI systems from different manufacturers. Our method enables generalisation with limited retraining data and our code is openly shared to enable further reproducibility and advancement of this field (as demonstrated in Section 4). A key contribution of our work is that human-AI complementarity is not always present (as was seen in 2 of 5 commercially-available TB systems) and in that setting our work shows that confidence-based deferral methods will not improve composite performance. In particular, the results from section 5 demonstrate the limitations of confidence based deferral strategies and are a useful tool to determine, given a particular dataset for training the deferral AI, whether one could expect to see any improvements from any confidence based deferral strategy. In real-world scenarios such an analysis could provide clear guidance on whether to use CoDoC.

For breast cancer screening in a large representative UK mammography dataset, CoDoC was superior in sensitivity to double-reading at the same specificity, and superior in sensitivity while maintaining specificity. “Double-reading” is regarded as the “gold standard” for performance in the UK and much of Europe [32 33 34 35]) never previously exceeded using AI[36 3738]. The same system maintained superior accuracy to both clinicians and the same diagnostic AI model even when the diagnostic AI was deployed out-of-distribution in a large US mammography dataset, only tuning the deferral AI on a small amount of out of distribution (OOD) data. Improvements in sensitivity and specificity were sustained for a wholly-separate diagnostic AI tool for US Mammography screening (from a different manufacturer) despite access to only 26 positive cancer cases for tuning.

CoDoC also conferred significant improvements in the resource-limited setting of TB screening in Bangladesh. CoDoC reduced the utilisation of Xpert tests for 3 of 5 commercially-available AI systems, by deciding when Xpert test utilisation should be decided by AI and when the decision should be deferred to a radiologist. This workflow has high real-world applicability as many TB screening centres using AI software already have the ability to route a subset of cases for radiology interpretation, while some countries specify that radiologists must be present at the time of CXR acquisition[39]. For 2 commercial AI systems, our CoDoC analysis demonstrated that confidence-based deferral would not improve performance over AI systems alone. In settings where radiologist interpretation is nevertheless considered mandatory for AI quality assurance, such CoDoC analysis might enable more cost-effective monitoring by highlighting situations in which radiologists performing quality-assurance of AI systems would be least likely to identify AI errors.

The breadth of clinical modalities demonstrates that CoDoC is highly clinically applicable because the deferral component is easily adaptable to multiple clinical workflows[39]. Even in one medical modality such as mammography, our results were robust in deferring to either single-reading or double-reading practice. We demonstrated that a variety of operating points could be chosen depending on the goals of the healthcare system, with statistically superior performance in clinically-applicable operating point regions. For example, a mammography centre might wish to optimise for either cancer detection rate or recall to assessment rate, and various CoDoC system configurations can be invoked depending on the balance between those goals with desired efficiencies for clinicians' time. Indeed our results suggest that deferral to a single reader might enable a screening programme to attain performance exceeding the gold standard of double-reading while only requiring a fraction of a single reader's time. Prospective and health economic outcomes studies will be required to confirm and quantify this potential benefit. The downstream effects of replacing a first reader (within an AI-enabled double-reader workflow) with CoDoC superior to the whole traditional double-reading workflow could also have a profound effect on the overall performance of AI-enabled double-reading. Future reader studies will be required to quantify this effect.

CoDoC performed well despite stress testing under multiple types of distribution shifts that commonly cause failures of medical AI in real-world settings. Particularly notable were results under two forms of shift that are common in the real world: shift in clinician performance and shift in population or site. It has been shown that clinicians' accuracy can vary significantly, both in terms of accuracy as well as in terms of the trade-off between sensitivity and specificity [32,40]. Reassuringly, CoDoC was able to generalise to multiple previously unseen readers in the UK mammography screening programme without any requirement for per-reader personalisation.

The variation in screening programmes between different hospitals or health systems is often sizeable[41] (with our experiments therefore exposing CoDoC to multiple shifts between health systems including changes in demographics, acquisition equipment, disease presentation, local clinical pathways). Despite significant differences from the diagnostic AI system's training data and an associated performance drop, the deferral AI was able to generalise to a previously unseen US hospital with minimal and realistic local training data needs. In particular, when we tuned the deferral AI using only 40 cases from a new population/site, CoDoC was able to improve upon the diagnostic accuracy of both the standalone AI and the clinician. In this setting, the deferral AI deferred a greater proportion of cases where diagnostic AI was less reliable than clinicians, suggesting that this paradigm could provide a valuable "safety net" for AI-enabled healthcare. This may enable local expert clinicians to mitigate concerns about failures of standalone diagnostic AI during deployment in new environments.

Comparison to relevant literature in AI

There is a long history of literature in machine learning that considers selective prediction systems that can refrain from making predictions on certain instances. This line of work traces back to the work of Chow et al[42], where the authors derived theoretically optimal algorithms in this setting. More recent reviews of this area can be found in Wiener et al[43]. Connections between selective prediction and active learning[44] have also been studied. These works differ from the deferral setting considered in this paper, since selective prediction ignores the accuracy of the human expert when the AI system abstains. The deferral setting was studied in Sontag et al[14] where the authors proposed a novel statistically consistent estimator for simultaneously learning a deferral model and the underlying prediction model. This was further extended to settings with multiple experts in subsequent work[45]. Optimising the performance of a human-AI team without restrictions on the deferral rate have been studied[46]. Other works have also proposed frameworks for AI models to defer to a domain expert in cases where the AI has low confidence in its inference [47] but require the ability to simultaneously learn both the classifier and deferral system. Others have proposed a model[48] to characterise human-AI (or human-human, AI-AI) complementarity, and demonstrated that complementarity may or may not exist in human-AI settings with the existence or degree of complementarity depending on a number of factors: the independence of human and AI decisions, existence of confidence scores for the predictions provided, and baseline individual performance of the human and AI. CoDoC extends and grounds these previous observations in the safety-critical medical AI domain, showing varying degrees of extractable complementarity between AI models and human experts, and proposing a reliable method for extracting it when available.

Many of the approaches above require co-training the deferral and diagnostic AI, which is not possible in medical AI settings where diagnostic classification tools are deployed in a "frozen" configuration by regulatory requirement and where access to the training pipeline for the diagnostic tools is not usual. Our work was inspired by this research, but we limited ourselves to deferral based on the confidence estimates of pre trained diagnostic AI models. This constraint for deferral systems to work with "black box" fixed diagnostic AI models also enables deferral to be studied in a wider variety of settings, since it absolves the requirement for access to the training pipeline and data for the diagnostic AI systems in each setting (which present significant practical hurdles to deferral paradigms that require co-training the deferral and diagnostic AI together). We found that our approach sufficed to obtain statistically significant improvements in performance with the CoDoC system, and that doing so decoupled the training of the deferral AI from the diagnostic AI which is highly advantageous in situations where the diagnostic AI is only available as a black box that cannot be modified (for example, due to IP or regulatory constraints). In future work, it would additionally be valuable to explore additional gains in diagnostic accuracy that could be obtained by co-training the diagnostic AI and deferral model, which might be possible for individual manufacturers in medical settings, even if not practicable in our setting of developing a single deferral wrapper for multiple different medical AI systems.

Limitations

In this study we evaluated performance under the assumption that clinicians and the diagnostic AI model perform independent case interpretation, as is approved in some clinical settings such as TB screening. However, in many settings clinicians use diagnostic AI models as an assistive tool, where prospective research will be required to establish the impact of CoDoC and where orthogonal work to CoDoC will be required to maximise its benefits. For example, it has been shown that the complementarity of AI tools for human experts is also dependent upon factors such as the operators' mental model, cognitive load, and trust[4], which could be optimised independently of the application of the CoDoC paradigm in a manner specific to each diagnostic AI tool. In particular, there is also evidence that providing AI decision-support can lead to systematic but unconscious biases on a clinician's decision-making process[4].

Our research demonstrated that improvements in accuracy were obtained using the CoDoC system while saving clinician's time compared to a standard AI-enabled workflow. The CoDoC framework already supports the introduction of tunable constraints/penalties on the deferral rate, and this could be adjusted based on desired savings for clinician time as a trade-off with composite accuracy. However, further health economic research and detailed per-hospital considerations would be needed to determine the right trade-offs, beyond the scope of the present work.

While our mammography test set was representative for UK practice, the US mammography dataset 2 was enriched for cancer prevalence compared to national practice. We simulated deployment scenarios for CoDoC with retrospective datasets, but quantifying the performance gains that result from clinician-AI interaction would require prospective reader studies and exploration of other aspects of human-AI complementarity orthogonal to the deferral decision - for example AI onboarding, trust and mental models.

The same limitation was also true of the CXR examinations used to triage Xpert tests for TB screening, in which multiple non-TB pathologies may be noticed by a radiologist but not classified by AI tools used to screen TB. Incorporating these tasks would require further research. Furthermore, although Xpert is regarded by WHO guidelines as an acceptable reference standard for evaluating AI systems, the gold standard would be full sputum culture for all participants. However, this was true for both AI and human radiologists in the dataset presented, so no selection or measurement bias was introduced and our approach was consistent with prior published work. CoDoC did not achieve uniform performance gains across the whole ROC curve. In the datasets we considered, the ROC range in which superiority was demonstrated coincided with regions of clinically-relevance (as illustrated by benchmarks of clinician sensitivity or specificity for screening decisions), but this may not be guaranteed for other applications of the CoDoC paradigm.

Beyond average diagnostic performance, variation among different population subgroups is an important concern as it can amplify health inequalities. This is a significant challenge for both standalone diagnostic AI systems and clinician experts, both shown to exhibit significant variation in population subgroup performance for a range of medical applications[49]. Preliminary analysis suggests that CoDoC does reduce variability in performance between different subpopulations, but further work is required to rigorously validate this, alongside further important distribution shifts for real-world medical AI, such as variations in instrumentation, acquisition and imaging technology.

Declarations

Acknowledgements

We would like to acknowledge multiple contributors to this international project: Stop TB Partnership hosted by UNOPS; Cancer Research UK, the OPTIMAM project team and staff at the Royal Surrey County Hospital who developed the UK mammography Optimam imaging database; our collaborators at Northwestern medicine and all members of the Etemadi Research Group for their continued support of this work; St. Clair Hospital, Klepchick BA, Andrus JM, Schaeffer RJ, and Sullivan JT; the authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

We would also like to thank L. Peng, D. Webster, U. Telang and D. Belgrave for their valuable feedback and support throughout the course of this project; D. Tran, N. de Freitas and K. Kavukcuoglu for critically reading the manuscript and providing feedback; R. Pilgrim, A. Kiani, and J. Rizk for work on partnership formation and engagement; R. May and E. Sutherland Robson for for assistance with project coordination; R. Najalfi and Z. Nabulsi for lung cancer domain expertise; S. Baur, S. Prabhakara for mammography domain expertise; M. Wilson for early engineering work.

Author Contributions

K.D., J.W., S.G., N.P., R.S., Y.B, P.K., T.C., and A.Karthikesalingam contributed to the conception and design of the study. J.W., S.M., S.S., M.S, T.S, G.C., and A.Karthikesalingam contributed to acquisition of the data; K.D., J.W., M.B., S.G., N.P., R.S., M.D., T.S., and T.C. contributed to analysis of the data; K.D., J.W., M.B., S.G., R.S., C.K., S.M., Z.Z.Q., J.C., K.G., J.Witowski, P.K., T.C., and A.Karthikesalingam contributed to interpretation of the data; K.D., J.W., S.G., M.B., N.P., R.S., S.A., L.C., M.D., J.F., A.Kiraly, T.K., S.M., B.M., V.N., S.S., M.S., and T.C. contributed to the creation of new software used in the work; K.D., J.W., M.B., S.G., R.S., P.M., Z.A., C.K., A.Kiraly, Z.Z.Q., J.C., K.G., J.Witowski, P.K., T.C., and A.Karthikesalingam contributed to drafting and revising the manuscript; K.D., J.W., M.B., S.G., R.S. P.M, Z.A., P.K., T.C., and A.Karthikesalingam contributed to paper organization and team logistics.

Competing Interests

This study was funded by Google LLC and/or a subsidiary thereof ('Google'). K.D., J.W., M.B., S.G., R.S., P.M., Z.A., S.A., Y.B., L.C., M.D., J.F., C.K., A.Kiraly, T.K., S.M., B.M., V.N., S.S., M.S., T.S., G.C., P.K., T.C., and A.Karthikesalingam are employees of Google and own stock as part of the standard compensation package. N.P. is a previous employee of Google and current employee of Microsoft. Z.Z.Q. and J.C. are employees of the Stop TB Partnership and collaborated with Google to support this research effort. K.G. and J.Witowski are employees of the NYU Grossman School of Medicine. K.G. and J.Witowski collaborated with Google to support this research effort.

Code Availability

Source code for the CoDoC system, trained deferral AI models and inference scripts will be made available under an open-source license at https://github.com/google-research/codoc.

The code is written in python 3 and consists of modules that implement the deferral AI. These accept a common data structure "DeferralData", consisting of 3 numbers per medical case: The average softmax score from an ensemble of diagnostic AI models, the opinion of a reader/clinician and the ground truth label. Given the tuning and evaluation set represented in the "DeferralData" data structure, these modules compute the deferral regions and deferral/prediction decisions on the evaluation sets. We will also release code to perform statistical hypothesis tests comparing two prediction mechanisms and establish statistical superiority or non-inferiority at a given significance level (p-value). These are released as python libraries without any dependence on proprietary code, only using open source libraries (numpy and scipy primarily).

Data Availability

The datasets from Northwestern Medicine, St. Clair Hospital, and NYU were used under licences for the current study, and are not publicly available. Datasets from the Stop TB Partnership and icddr,b were used under a licence for the current study, and are not publicly available. Applications for access to the OPTIMAM database can be made at https://medphys.royalsurrey.nhs.uk/omidb/getting-access/. NLST: https://biometry.nci.nih.gov/cdas/learn/nlst/images/. We will make available the data necessary to reproduce our results (model scores from diagnostic AI, clinician opinions and ground truth labels) for the tuning/test sets from all datasets used in the paper, at https://github.com/google-research/codoc.

References

1. Liu, X. *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* **1**, e271–e297 (2019).

2. Aggarwal, R. *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* **4**, 65 (2021).

3. Center for Devices & Radiological Health. Artificial Intelligence and Machine Learning (AI/ML) Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (2021).

4. Gaube, S. *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* **4**, 1–8 (2021).

5. Kim, H.-E. *et al.* Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* **2**, e138–e148 (2020).

6. Rajpurkar, P. *et al.* CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digit Med* **3**, 115 (2020).

7. Tschandl, P. *et al.* Human-computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).

8. Chi, E. A. *et al.* Development and Validation of an Artificial Intelligence System to Optimize Clinician Review of Patient Records. *JAMA Netw Open* **4**, e2117391 (2021).

9. Ruamviboonsuk, P. *et al.* Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Medicine* **2**, 1–9 (2019).

10. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).

11. Lee, C. S. & Lee, A. Y. Clinical applications of continual learning machine learning. *The Lancet. Digital health* **2**, (2020).

12. DEFINE_ME. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00076-5/fulltext.

13. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Continual learning in medical devices: FDA's action plan and beyond. *The Lancet. Digital health* **3**, (2021).

14. Mozannar, H. & Sontag, D. Consistent Estimators for Learning to Defer to an Expert. (2020).

15. Hendrycks, D. & Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. (2016).

16. Leibig, C. *et al.* Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *Lancet Digit Health* **4**, e507–e519 (2022).

17. D'Amour, A. *et al.* Underspecification Presents Challenges for Credibility in Modern Machine Learning. (2020) doi:10.48550/arXiv.2011.03395.

18. Programme, G. T. Use of Xpert MTB/RIF and Xpert MTB/RIF Ultra on GeneXpert 10-colour instruments: WHO policy statement. https://www.who.int/publications/i/item/9789240040090 (2021).

19. Kernel density estimation. https://en.wikipedia.org/wiki/Kernel_density_estimation (2005).

20. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).

21. Mustafa, B. *et al.* Supervised Transfer Learning at Scale for Medical Imaging. (2021).

22. Azizi, S. *et al.* Big Self-Supervised Models Advance Medical Image Classification. (2021).

23. [No title]. https://arxiv.org/pdf/2108.04800.pdf.

24. Oakden-Rayner, L. & Palmer, L. Docs are ROCs: A simple off-the-shelf approach for estimating average human performance in diagnostic studies. (2020).

25. Stadnick, B. *et al.* Meta-repository of screening mammography classifiers. (2021) doi:10.48550/arXiv.2108.04800.

26. Calibrating computer-aided detection (CAD) for TB. https://tdr.who.int/activities/calibrating-computer-aided-detection-for-tb.

27. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. *The Lancet Digital Health* **3**, e543–e554 (2021).

28. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).

29. [No title]. https://arxiv.org/pdf/1711.11279.pdf.

30. Kim, B. *et al.* Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). (2017).

31. Kompa, B., Snoek, J. & Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* **4**, 4 (2021).

32. Freeman, K. *et al.* Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* **374**, n1872 (2021).

33. [No title]. https://www.rcr.ac.uk/system/files/publication/field_publication_files/bfcr199-guidance-on-screening-and-symptomatic-breast-imaging.pdf.

34. Use of double reading in mammography screening. https://healthcare-quality.jrc.ec.europa.eu/european-breast-cancer-guidelines/organisation-of-screening-programme/how-mammography-should-be-read (2019).

35. Breast screening: quality assurance standards in radiology. *GOV.UK* https://www.gov.uk/government/publications/breast-screening-quality-assurance-standards-in-radiology (2011).

36. [No title]. https://www.medrxiv.org/content/10.1101/2021.02.26.21252537v1.full.pdf.

37. The potential of AI to replace a first reader in a double reading breast cancer screening program: a feasibility study. *Screen Point* https://screenpoint-medical.com/evidence/the-potential-of-ai-to-replace-a-first-reader-in-a-double-reading-breast-cancer-screening-program-a-feasibility-study/ (2021).

38. Larsen, M. *et al.* Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program. *Radiology* (2022) doi:10.1148/radiol.212381.

39. Qin, Z. Z. *et al.* Early user experience and lessons learned using ultra-portable digital X-ray with computer-aided detection (DXR-CAD) products: A qualitative study from the perspective of healthcare providers. *Arxiv (Forthcoming)* (2022).

40. Barlow, W. E. *et al.* Accuracy of screening mammography interpretation by characteristics of radiologists. *J. Natl. Cancer Inst.* **96**, 1840–1850 (2004).

41. Demchig, D. *et al.* Observer Variability in Breast Cancer Diagnosis between Countries with and without Breast Screening. *Acad. Radiol.* **26**, 62–68 (2019).

42. On optimum recognition error and reject tradeoff. https://ieeexplore.ieee.org/abstract/document/1054406/.

43. Wiener, Y. & Ṭekhniyon, M. Ṭekhnologi L.-Y. F. L.-M. H.-M. *Theoretical Foundations of Selective Prediction*. (2013).

44. Gelbhart, R. & El-Yaniv, R. The Relationship Between Agnostic Selective Classification, Active Learning and the Disagreement Coefficient. *J. Mach. Learn. Res.* **20**, 1–38 (2019).

45. Keswani, V., Lease, M. & Kenthapadi, K. Towards Unbiased and Accurate Deferral to Multiple Experts. (2021).

46. Wilder, B., Horvitz, E. & Kamar, E. Learning to Complement Humans. (2020) doi:10.48550/arXiv.2005.00582.

47. Vijay Keswani Yale University, New Haven, CT, USA, Matthew Lease University of Texas at Austin, Austin, TX, USA & Krishnaram Kenthapadi Amazon AWS AI, East Palo Alto, CA, USA. Towards Unbiased and Accurate Deferral to Multiple Experts. *ACM Conferences* https://dl.acm.org/doi/10.1145/3461702.3462516.

48. Website. https://www.pnas.org/doi/10.1073/pnas.2111547119.

49. Chen, I. Y. *et al.* Ethical Machine Learning in Healthcare. *Annu Rev Biomed Data Sci* **4**, 123–144 (2021).

Figures

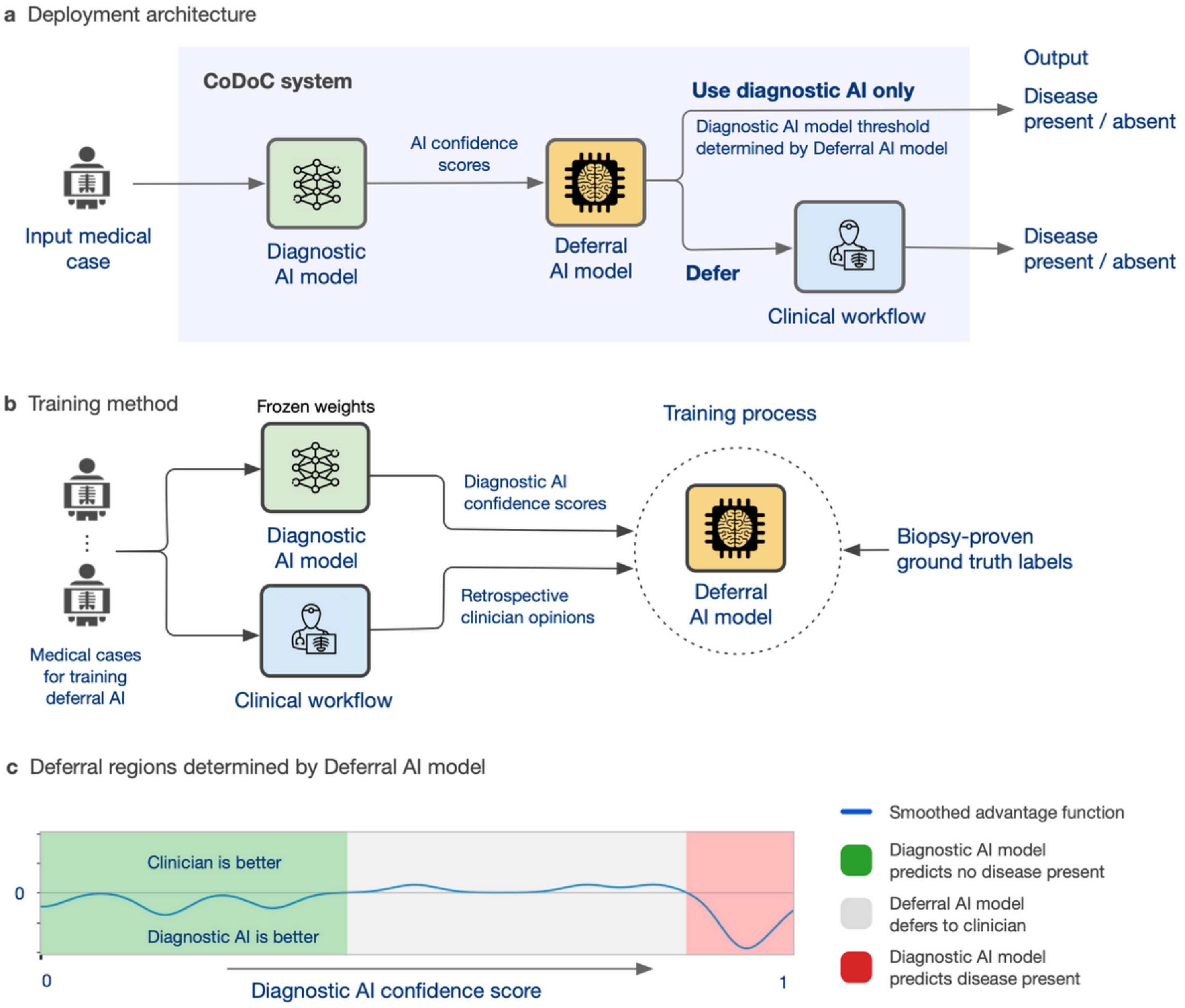
a Deployment architecture
CoDoC system
Input medical case
Diagnostic AI model
AI confidence scores
Deferral AI model
Use diagnostic AI only
Diagnostic AI model threshold determined by Deferral AI model
Defer
Clinical workflow
Output
Disease present / absent
Disease present / absent
b Training method
Frozen weights
Medical cases for training deferral AI
Diagnostic AI model
Clinical workflow
Diagnostic AI confidence scores
Retrospective clinician opinions
Training process
Deferral AI model
Biopsy-proven ground truth labels
c Deferral regions determined by Deferral AI model
Clinician is better
Diagnostic AI is better
0
0
1
Diagnostic AI confidence score
Smoothed advantage function
Diagnostic AI model predicts no disease present
Deferral AI model defers to clinician
Diagnostic AI model predicts disease present

Figure 1

Figure 1 a-c: ***CoDoC training and deployment architecture. a.*** *We obtain the confidence score (specifically the softmax score output by a deep learning model) output from the diagnostic AI model for each image. These are then fed into a deferral AI model, which decides to either use the confidence score of the diagnostic model and apply a threshold to it to make a final disease/no disease prediction or defer, in which case the medical image is diagnosed using a standard clinical workflow involving one or more clinicians. We refer to the composite decision-making apparatus comprising the deferral AI and the selectively invoked diagnostic AI model or non-AI clinical workflow as the CoDoC system.* ***b****. The diagnostic AI model is pretrained on an initial training dataset (not shown, and which need not be accessible when the deferral AI model is being trained). To create the training data for the deferral AI model, a fresh set of medical images (referred to in this paper as the tuning data) are passed through the diagnostic AI models. Clinician opinions and ground truth labels are also collected for the tuning data, to obtain a tuple (AI scores, clinician opinions, ground truth label) for each medical image in the tuning data. These tuples are used to train the deferral AI to decidewhen to defer, and if not deferring, to choose an operating point for the diagnostic AI to make a final diagnostic prediction in order to maximise the accuracy (in terms of either sensitivity or specificity) of the deferral AI system.* ***c.*** *Using the AI and clinician accuracy, summarised in the relative accuracy graph shown, the deferral AI system learns deferral regions which are depicted in grey. When the average softmax score from the diagnostic AI models lies in one of grey regions, CoDoC defers to a clinical workflow and presents the outcome of the clinical workflow as the final diagnosis. When the confidence score lies in the green region, CoDoC predicts that the disease is absent, and when the confidence score lies in the red region, CoDoC predicts that the disease is present.*

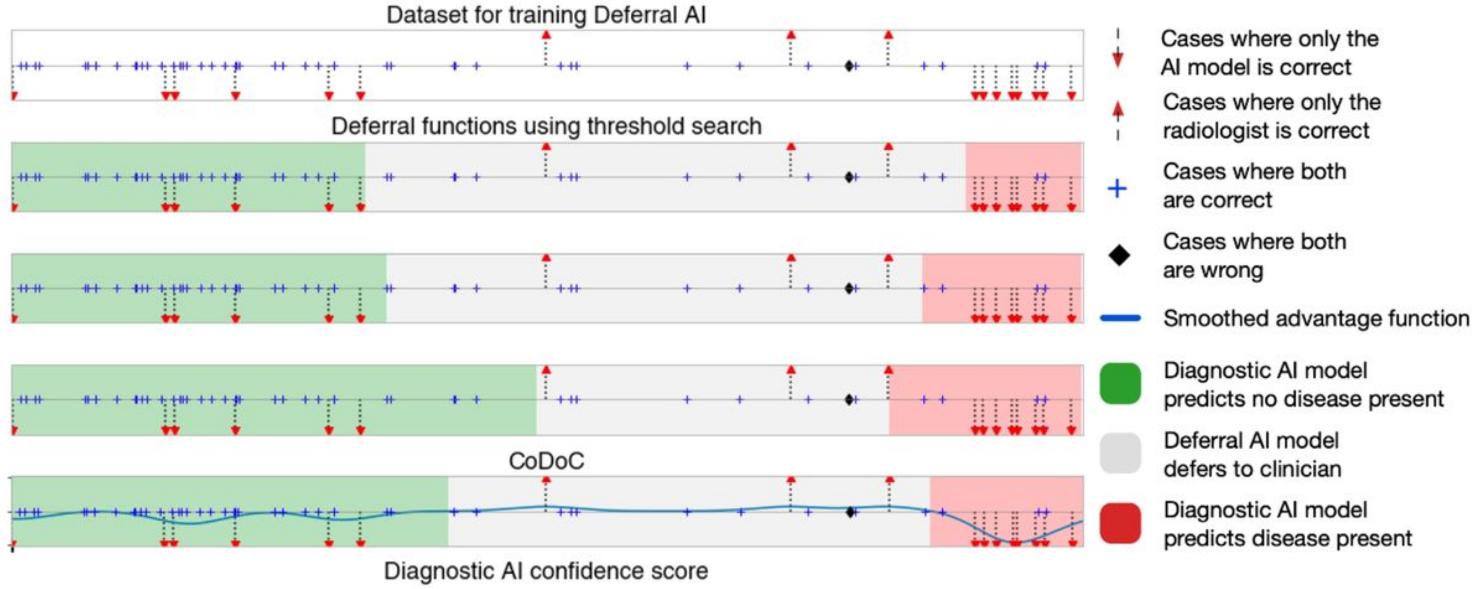

Figure 2

Figure 1d | Underspecification in Threshold Search. *The dataset is plotted by tagging each datapoint with a separate marker depending on whether the AI model makes the correct prediction and whether the human clinician makes the correct prediction for that case. Examples where only a human makes the* *correct prediction are advantageous to defer on (indicated by the upward arrows) and cases where only the model makes the correct prediction are indicated by downward arrows. The top three figures plot various optimal solutions for threshold search all of which obtain the exact same performance on the dataset. How to choose between these options is unclear, and the naive threshold search does not offer any guidance on the same. In our approach, we first compute a smoothed representation of the discrete advantage function and the zero crossings of this smoothed advantage function. The parameters of the smoothing algorithm are tuned based on the performance of the selected thresholds on a validation set.*

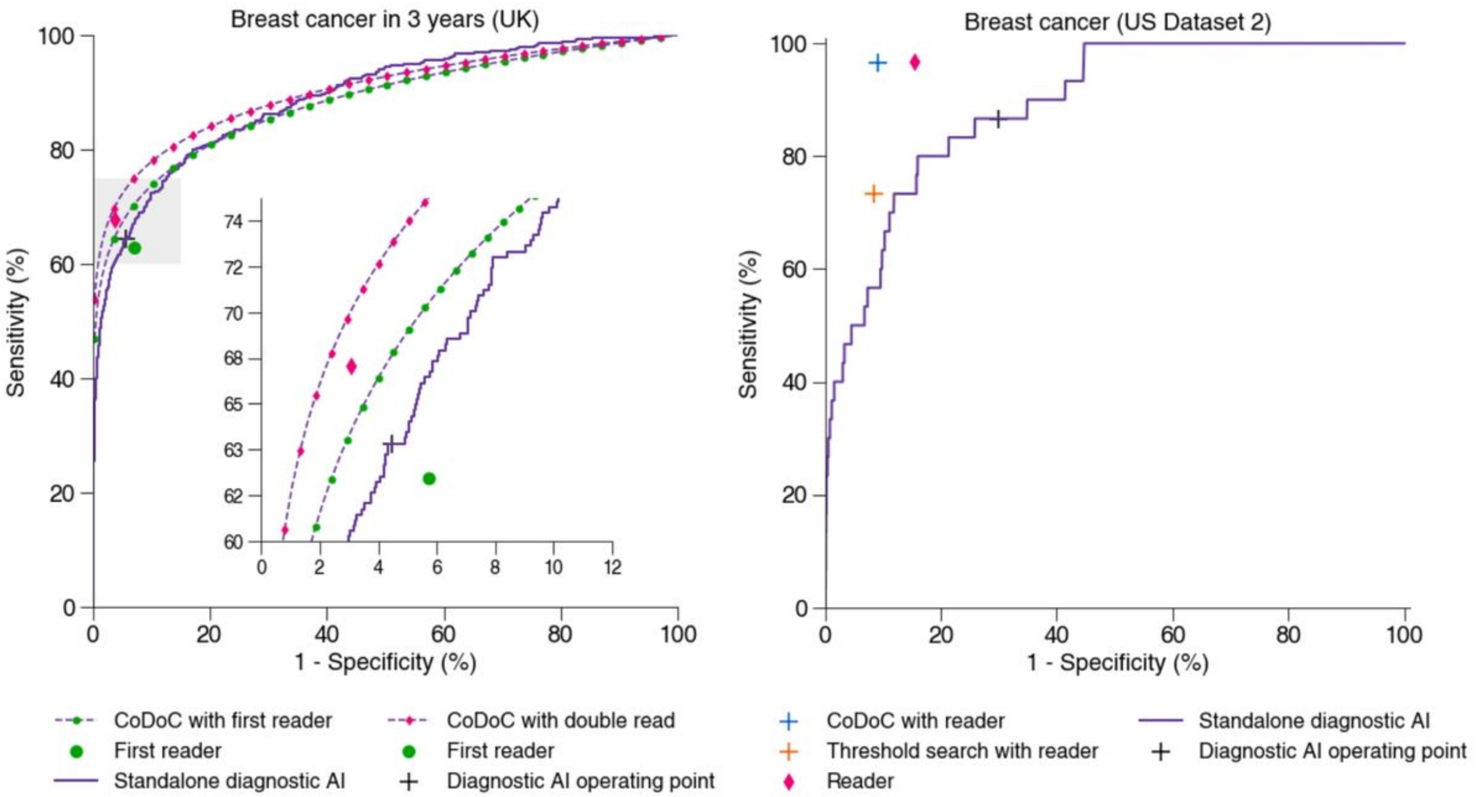

Figure 3

Figure 2 | Performance of CoDoC in breast cancer prediction compared to a standalone diagnostic AI system and clinical readers. *Defer to the single read clinical workflow and defer to double-reader with arbitration clinical workflow in the UK (Mammography Diagnostic AI 1, UK Mammography Dataset) (Left). Defer to the single read clinical workflow in the US (Mammography Diagnostic AI 2, US Mammography Dataset 2) (Right). Several final operating points could be chosen for the output of CoDoC. We illustrate the Pareto frontier of such optimal operating points by computing a summary ROC (sROC) curve*[24] *to enable easier visualisation. We note that we only include CoDoC models for which we can obtain statistically significant improvements on sensitivity or specificity compared to the diagnostic AI and the clinical workflow. Thus, on some datasets (like the US Dataset 2), we only show a single CoDoC model that improves specificity while remaining non-inferior on sensitivity. We were unable to improve sensitivity on this dataset as the clinical workflow already achieved a very high sensitivity.*

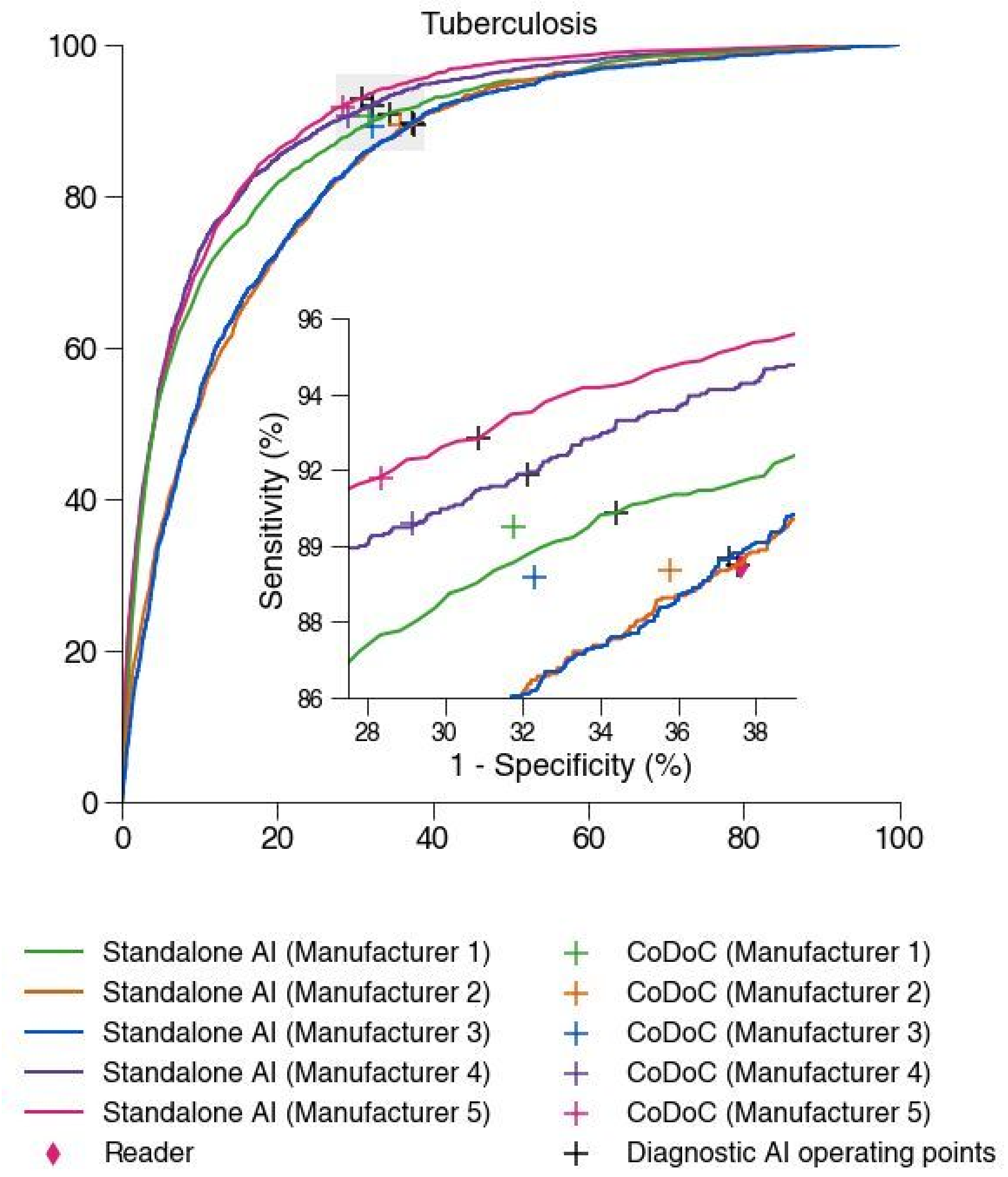

Figure 4

Figure 3: Performance of CoDoC in tuberculosis prediction. Performance of CoDoC in TB screening from CXRs compared to 5 standalone AI systems (Standalone AI Manufacturers 1-5), and compared to the performance of radiologists (Reader). We illustrate the ROC curve for each standalone AI system. For each standalone AI system we highlight the performance of CoDoC (when using CoDoC to defer to the Reader), displaying a clinically-applicable operating point with CoDoC optimising specificity.

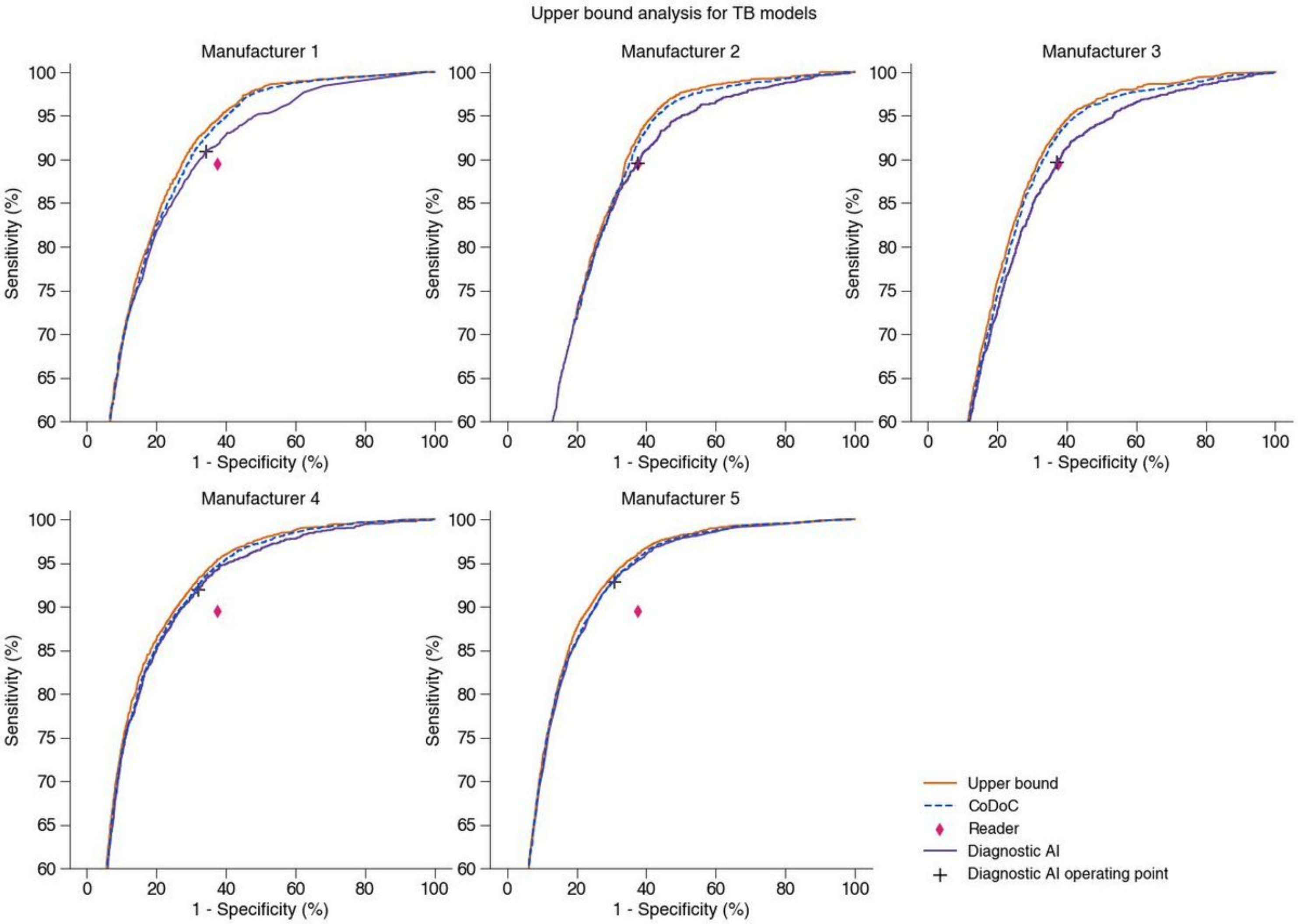

Figure 5

Figure 4: Limitations of confidence-based deferral. *The bottom row depicts the strongest performing commercially available TB models where the potential improvement from a confidence based deferral strategy (observable as the gap between the orange curve and the purple curve) is very small. However, on the top row, we see the same for the other three commercially available TB models that there is a margin for improvement via confidence based deferral, and the CoDoC system is indeed able to achieve improvements in this setting as demonstrated in Figure 3.*

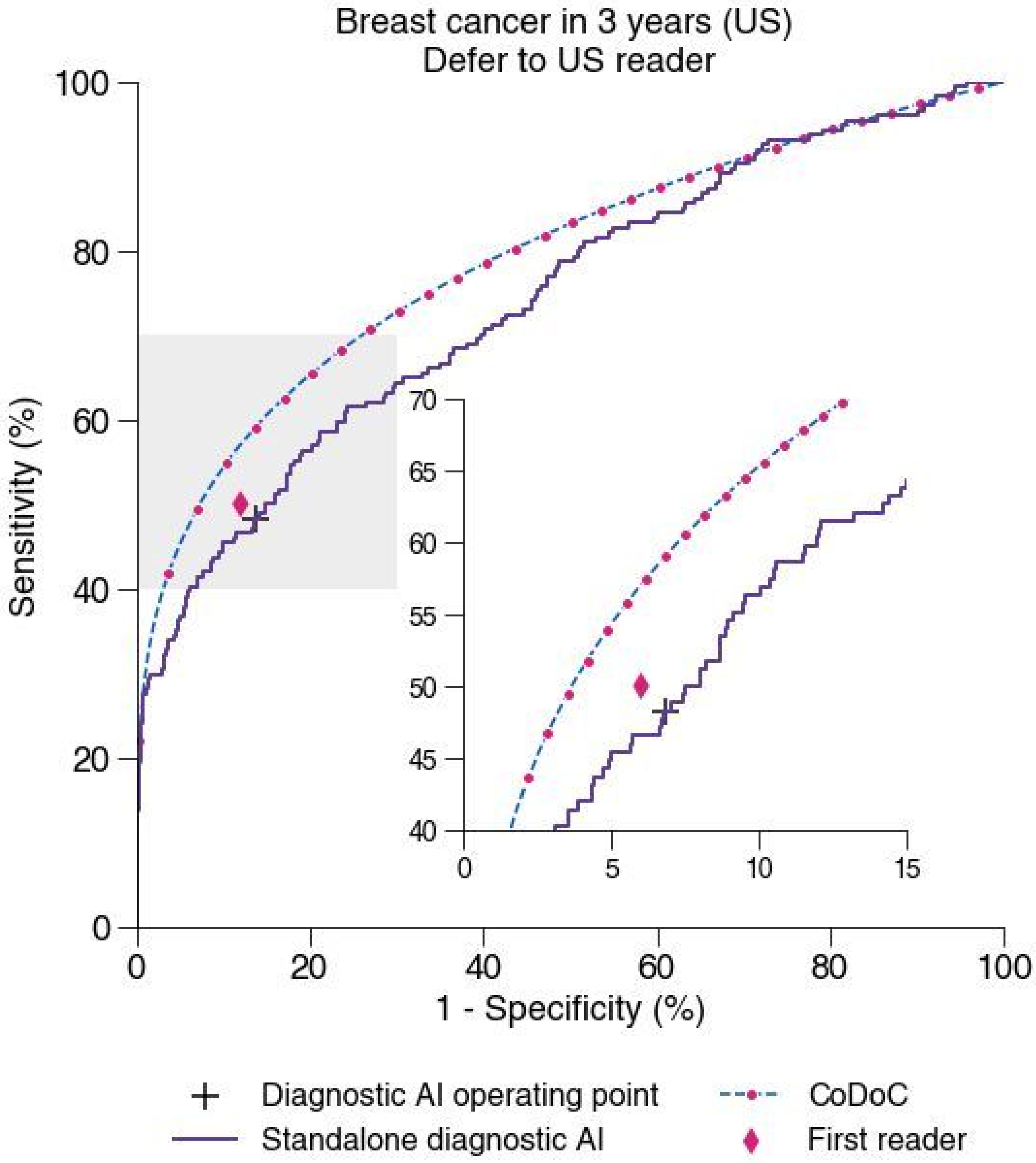

Figure 6

Figure 5: Performance of CoDoC in breast cancer prediction on a US mammography dataset (US Mammography Dataset 2) *In this screening centre, historical images were read by a single reader, who had access to both mammography and digital breast tomosynthesis. While the diagnostic AI model had not previously been trained on this population, CoDoC had been tuned on a held-out subset.*

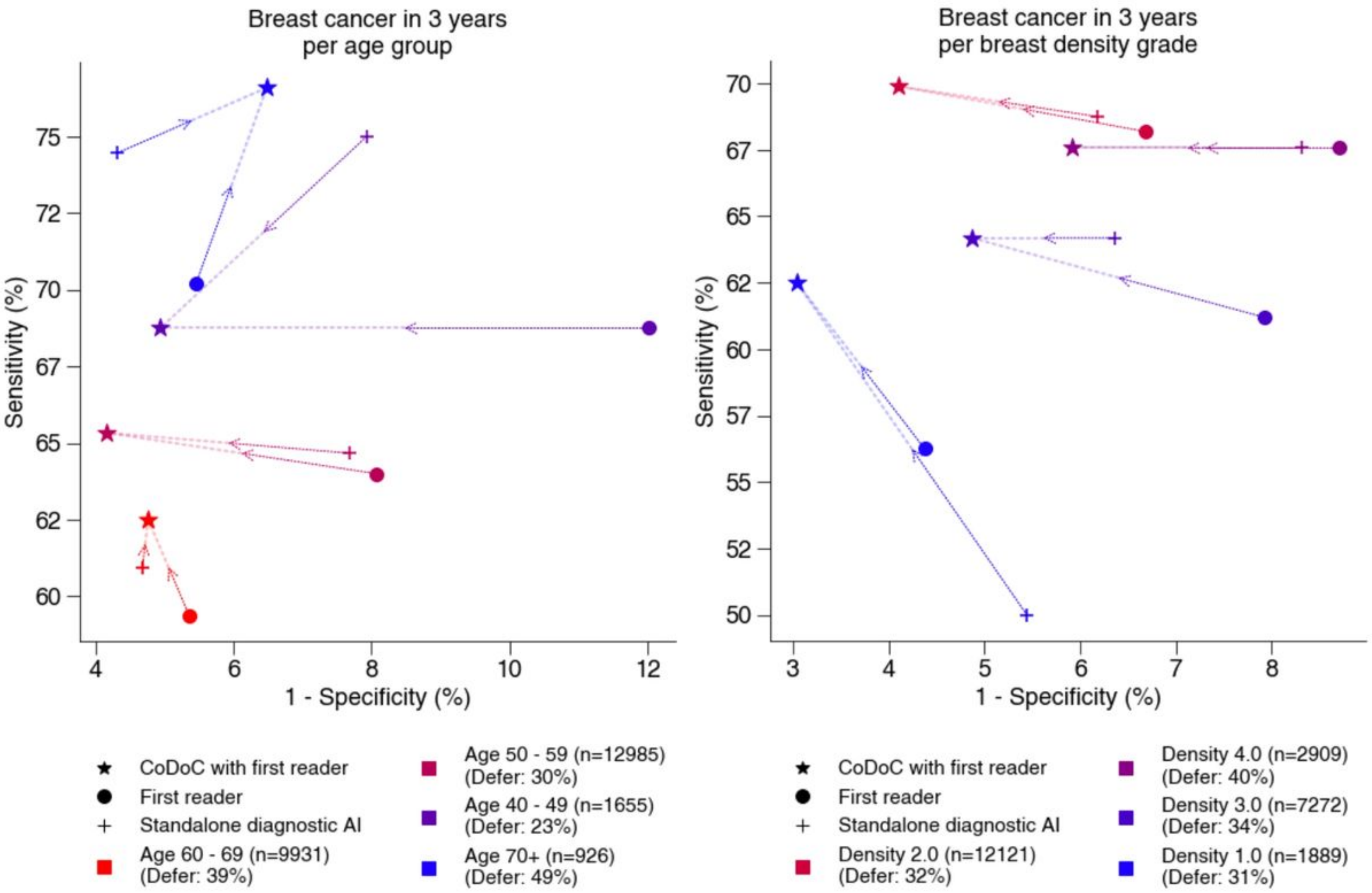

Figure 7

Figure 6: Subgroup breakdown across age groups (left) and breast density grades (right) on the UK dataset. *Across subgroups, CoDoC ( ) combines performance from both the clinician (●) and standalone diagnostic AI system (+). CoDoC shown is "Defer to First reader OP 1" from Table 1.*

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CoDoCSupplementaryInformationgocodocappendix.docx