

Assistance curve analysis: evaluating the clinical value of AI-assisted decision making

Marwan Alqumbaey

marwanalqumbaey@gmail.com

Sana'a University

Yousef M. Nasher

Sana'a University

Method Article

Keywords: decision curve analysis, net benefit, clinical decision support, artificial intelligence, reader study, clinical usefulness

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10899895/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background.

Clinical AI is almost never used alone. It shows something to a doctor, and the doctor decides. Yet the measures used to judge them describe the model, not the decision. A 2025 overview of 32 performance measures for clinical AI covers five domains, and current standard metrics do not explicitly evaluate the decisions of an assisted clinician in a paired setting.

Methods.

We define the net benefit of assistance, $\Delta\text{NB}(p)$, as the difference in net benefit between the decisions a clinician makes with an AI available and those they make without it. We show that ΔNB depends only on cases where the clinician changed their mind, and give a closed form in four quantities: how often the AI moves the clinician each way, and how often it is right when it does. This yields a break-even rule, a number needed to consult, and an override profile. We checked the estimators by simulation and applied the method to Collab-CXR, a public study of 227 radiologists reading chest X-rays with and without a CheXpert-trained model.

Results.

$\Delta\text{NB}(p) = [r_U(\pi_U - p) + r_L(p - \pi_L)]/(1 - p)$. Across 27,720 paired decisions by 33 radiologists, the AI made them less accurate by 0.83 percentage points while making their decisions better. ΔNB was + 0.0022 at $p = 0.05$ (95% CI + 0.0007 to + 0.0038), a 7% gain on the unaided radiologist's net benefit, or one extra net true positive per 455 decisions. The benefit was clear between thresholds of 0.02 and 0.05; above 0.05 the interval included zero, and below about 0.019 both arms were beaten by acting on every finding. The break-even threshold was 0.201. Two further designs covering 194 radiologists agreed in direction. Effects ranged from + 0.0131 for atelectasis to - 0.0024 for cardiomeastinal abnormality.

Conclusions.

The value of an AI is made on the decisions it changes and cannot be read off accuracy. Judging by accuracy would have rejected a tool that improved decisions. The method needs only what a paired reader study already collects, and free software is provided for R and Python.

Background

An AI that reads a chest X-ray does not admit the patient. A radiologist does, after seeing what the AI said. The thing with consequences is not the model. It is the decision a doctor makes while looking at

the model's output, and everything that happens to the patient follows from that decision rather than from the model's area under the curve.

The last five years have taught us a great deal about what happens at that moment, and almost none of it is comfortable. When the AI is right, doctors get better. When it is wrong, they get worse. Tschandl and colleagues showed this for skin cancer: correct support raised rater accuracy by a median of 9.5 percentage points, faulty support lowered it by a median of 6.3, and experts were misled as well as novices [1]. Gaube and colleagues gave 265 physicians chest X-rays with advice that was sometimes wrong, labelled as coming either from an AI or from a radiologist. Accuracy fell whenever the advice was bad, and the label made no difference to whether it was followed, though the radiologists rated the AI-labelled advice lower in quality while following it just as readily [2]. Kücking and colleagues recorded 1,338 decisions by 223 nurses and physicians assessing wound images: accuracy was 92.1% when the recommendation was correct and 55.6% when it was wrong, against 76.5% unaided [3]. Jabbour and colleagues randomised 572 clinicians, of whom 457 completed at least one vignette and were analysed, and found that a systematically biased model cut diagnostic accuracy by 11.3 percentage points (95% CI 7.2 to 15.5), while image-based explanations recovered only a non-significant 2.3 points [4].

Two findings from this literature matter more than the rest for what follows.

The first is that averages hide the effect. Kiani and colleagues gave eleven pathologists a deep learning assistant for liver cancer subtyping and found no significant change in mean accuracy ($p = 0.184$), although accuracy did improve in a sensitivity analysis restricted to the nine pathologists of defined experience level ($p = 0.045$). Underneath either average, assistance helped when the model was correct (OR 4.29) and hurt when it was not (OR 0.25) [5]. Yu and colleagues found the same heterogeneity across 140 radiologists and 15 chest X-ray tasks, and reported that neither years of experience nor familiarity with AI predicted who would benefit [6]. That analysis uses designs 1 and 2 of the experiment we reanalyse here, so it is not independent of our applied results. A single summary number for the effect of AI assistance is averaging over something that changes sign.

The second is that the harm is not confined to the moment of use. Dratsch and colleagues documented automation bias in mammography, where an incorrect suggestion from a purported AI system cut the share of correctly rated mammograms from about 80 percent to between 20 and 46 percent, with the most experienced readers more resistant at 45.5% [7]. Budzyń and colleagues compared the three months before and after AI was introduced into routine colonoscopy at four centres, and found that the adenoma detection rate of unassisted procedures fell from 28.4% to 22.4%, which they read as a possible deskilling effect in an observational design [8].

So the field knows that assistance can help and can harm, and that the balance depends on cases nobody is counting separately. What it does not have is a way to measure that balance in units that mean something clinically.

The measures in current use were built for something else. Discrimination and calibration describe what a model predicts. Accuracy, sensitivity and specificity describe a classifier. Van Calster and colleagues assessed 32 performance measures for predictive AI in medicine across five domains, namely discrimination, calibration, overall performance, classification and clinical utility, and set out a reporting minimum [9]. On our reading, every measure in that overview evaluates the model, and none evaluates the assisted decision.

Where the assisted decision is evaluated, it is usually by comparing accuracy with and without help, or by one of the reliance measures developed in human-computer interaction research. Schemmer and colleagues define relative AI reliance as the share of cases where a clinician who was wrong adopts correct advice, and relative self-reliance as the share where a clinician who was right resists incorrect advice [10]. Those tell you something real about behaviour, but they are plain counts, so a missed pneumothorax and an unnecessary follow-up film sit on the same scale. In medicine the two are almost never equal, and the balance between them is exactly what decides whether a decision aid is worth having.

Decision analysis solved this problem for prediction models a long time ago. Pauker and Kassirer made the trade-off explicit through the threshold probability, the risk at which treating and withholding treatment look equally good to the clinician [11], and later derived the thresholds at which testing is worthwhile [12]. Vickers and Elkin turned that into decision curve analysis, which reports the clinical value of acting on a model across the range of thresholds a clinician might hold [13, 14]. The method has since been extended to choices between several treatments [15], to Bayesian inference [16], and to questions about sampling uncertainty [17, 18]. The closest work in the assisted setting reports decision-analytic summaries alongside accuracy rather than in place of it. Sunwoo and colleagues, in a multi-reader study of AI-assisted stroke triage, quantified clinical utility with a net reclassification improvement, a number needed to screen and a benefit-to-harm ratio [19]. Those quantities do price the two errors against each other, which plain accuracy does not, but each is a single number computed at one operating point: none is a function of the threshold probability, so none shows where along the range of clinical questions the sign of the effect changes, and none is defined on the cases the aid actually altered. Comparing two strategies by their net-benefit difference is standard [20], but the methodological contribution of this work lies in the switch decomposition. This study provides a formal extension of decision curve analysis to the situation that now dominates clinical AI: a doctor and a model deciding together.

This paper makes that extension. We define the net benefit of assistance, show that it comes entirely from the cases where the clinician changed their mind, and reduce it to a form built from four numbers a clinician can hold in their head. We then apply it to a public study of 227 radiologists and find that accuracy and clinical value moved in opposite directions.

Methods

Setting and notation

Take n cases, numbered $i = 1$ to n . Let d_i be 1 if the condition is present and 0 if it is not. There is one action available, and the clinician either takes it or does not: biopsy, admit, treat, follow up. For each case we need three things: c_i , what the clinician decided without the AI; y_i , what the same clinician decided with the AI; and d_i , what was actually true. One more is useful where it exists: s_i , the AI's predicted probability, which the override profile uses. Both c and y must come from the same cases, which is the standard paired reader design.

One assumption should be said at the start. Net benefit is normally applied to a rule of the form “act if the predicted risk is above p ”. Here c and y are decisions made at a threshold we never see, so evaluating them at p asks what those decisions were worth to a clinician who holds threshold p . This is the same step decision curve analysis takes whenever it evaluates a fixed clinical policy rather than a model score, and the sensitivity analysis using the radiologists' own stated probabilities, where the action rule is fixed at the threshold by construction, is the check on it.

Net benefit and the net benefit of assistance

For a threshold probability p between 0 and 1, the net benefit of any set of decisions y is [13]

$$NB(y; p) = TP/n - w \times FP/n, \text{ where } w = p/(1 - p) \quad (1)$$

TP and FP are the counts of true and false positive decisions. The threshold probability is the risk at which acting and not acting look equally good to the clinician, and w is the price of an unnecessary action in units of a necessary one. Net benefit is measured in true positives per case.

We define the net benefit of assistance as

$$\Delta NB(p) = NB(y; p) - NB(c; p) \quad (2)$$

If $\Delta NB(p)$ is positive, then at that trade-off, showing the AI to the clinician did more good than harm. Since the trade-off is not the same for every clinical question, we plot ΔNB across the range of p that fits the question, the same way an ordinary decision curve is plotted. We call that plot the assistance curve.

Where the value comes from

Result 1. $\Delta NB(p)$ depends only on the cases where the clinician changed their decision.

Net benefit is linear in the decisions: $NB(y; p) = (1/n) \sum_i y_i [d_i - w(1 - d_i)]$. Subtracting the unaided from the assisted gives

$$\Delta NB(p) = (1/n) \sum_i (y_i - c_i) [d_i - w(1 - d_i)] \quad (3)$$

Every case where y_i equals c_i drops out.

The proof is short. What it says is not. An AI's behaviour on the great majority of cases, the ones where the clinician was going to do the same thing either way, contributes nothing at all to its clinical value. Accuracy is measured across all cases. Value is made on a small minority of them. Two systems with the same discrimination can have opposite clinical value, because they change different decisions.

Now split the changed cases in two. Let U be the cases where $c_i = 0$ and $y_i = 1$, in which the AI moved the clinician to act, and L the cases where $c_i = 1$ and $y_i = 0$, in which it moved the clinician to stand down. Write r_U and r_L for how often each happens as a share of all cases, and π_U and π_L for how often the condition was really present within each group.

Result 2. Splitting (3) over U and L , the contribution from U is $r_U[\pi_U - w(1 - \pi_U)]$, which over a common denominator is $r_U(\pi_U - p)/(1 - p)$. The contribution from L is $r_L[w(1 - \pi_L) - \pi_L] = r_L(p - \pi_L)/(1 - p)$. So

$$\Delta NB(p) = [r_U(\pi_U - p) + r_L(p - \pi_L)] / (1 - p) \quad (4)$$

Because c and y are fixed decision vectors, ΔNB is affine in $w = p/(1 - p)$: $\Delta NB = (u_1 - l_1)/n + w(l_0 - u_0)/n$, where u_1, u_0, l_1, l_0 are the four switch counts. The assistance curve therefore has two degrees of freedom and is a reparameterisation of the switch table rather than a curve whose decision set changes with p , as in ordinary decision curve analysis. This is what makes the four-number account in Eq. (4) complete, and it is also why the probability-based analysis is reported alongside.

Equation (4) reduces the whole question to four numbers: how often the aid moves you in each direction, and how often it is right when it does.

The break-even rule. Moving up is worth it only when π_U is greater than p . Moving down is worth it only when π_L is less than p . In plain words, an AI is worth following upwards only if more than p of the patients it talks you into treating really have the disease, and worth following downwards only if fewer than p of the patients it talks you out of treating really have it. Both numbers can be measured in any paired reader study, and neither needs the model's probabilities to be calibrated, or even to exist.

The break-even threshold. When r_U and r_L differ, the numerator of (4) is linear in p and the denominator is positive, so ΔNB changes sign at most once, at

$$p^* = (r_L \pi_L - r_U \pi_U) / (r_L - r_U) \quad (5)$$

The crossing means something only when p^* falls inside $(0, 1)$. Outside that range, and when r_U equals r_L , ΔNB keeps one sign at every threshold a clinician could hold. Where r_U is the larger, ΔNB is positive below p^* and negative above it, and the reverse where r_L is larger. Reporting p^* tells a reader straight away which clinical questions the tool is safe for.

Which study designs allow which analyses

ΔNB itself can be estimated wherever decisions are observed with and without the aid, including a design that never reads the same case twice, because Eq. (2) is only a difference between two population net benefits.

The split into up-switches and down-switches cannot. The quantities r_U , r_L , π_U and π_L are defined on case-level changes of mind, and an unpaired design never sees those. So a study that wants to know why an AI helps, and not only whether it helps, has to be paired. That is a design recommendation that falls out of the theory rather than being added from outside.

Paired designs bring their own problem, which is that the clinician sees each case twice and may remember the first reading. The usual protections apply: a washout gap, randomised order of the conditions, and blinding of the first decision. Where they are missing, any order effect lines up with the AI condition, and the estimate cannot be put down to the aid alone.

What to report

We report $100 \times \Delta\text{NB}(p)$ as the net gain per 100 decisions, read as the number of extra correct decisions to act after paying for the false positives at the price the clinical question sets. Net benefit values are small numbers, so we also give the relative gain, $\Delta\text{NB}(p)$ divided by the unaided clinician's net benefit at the same p , which says what share of the clinician's own clinical value the aid added or removed. The number needed to consult, $\text{NNC} = 1/\Delta\text{NB}(p)$, is the number of cases that must be shown to the AI to gain one extra net true positive. This is the test trade-off of decision curve analysis [21], applied to the difference between the assisted and unassisted clinician rather than to the addition of a marker; it swings wildly when ΔNB is near zero, so we report it only where the interval is clear of zero. The switch rate, $r_U + r_L$, tells you how much the AI influences behaviour, and ΔNB divided by it tells you what the average changed decision was worth. Separating the two distinguishes an aid that is usually right but rarely listened to from one that is always listened to and usually wrong.

Net benefit adds up across separate groups of cases, so you can work out what any group contributed to ΔNB . Grouping cases by the AI's own score gives the override profile: the range of outputs where showing the model improved decisions and the range where it did not. That turns a verdict into a display rule. The profile depends on the threshold, so we report it at the primary one.

One caution goes with all of these. An assistance curve compares a clinician with themselves, so it can be positive in a region where both arms are worse than acting on everyone or on no one. The ordinary decision curve, with treat-all and treat-none drawn on it, must be reported alongside, and the assistance result read only where the clinician beats both.

Uncertainty

Vickers and colleagues have argued against significance testing for net benefit [17]. When a decision has to be made, you take the option with the higher expected value whatever the p value says, and uncertainty is better thought of as the value of doing more research. We follow that, and report percentile bootstrap intervals together with the bootstrap probability that ΔNB is above zero at each threshold.

Resampling has to respect the structure of the data. Decisions made by the same radiologist are not independent, and neither are the 14 findings scored on one patient's image. We ran the bootstrap four ways, clustered on the radiologist, on the patient-finding, on the patient, and unclustered, and report the radiologist-clustered interval, which is the widest at every threshold on which the conclusion rests. The primary analysis rests on 33 radiologists, and percentile intervals over that few clusters can be anti-conservative, so we also computed a cluster sign-flip permutation distribution, in which each radiologist's contribution to ΔNB has its sign flipped independently over 100,000 sign vectors, as a check on the intervals. All intervals come from the Python implementation. The four clustering schemes and the design-level estimates use 2,000 replicates, the finding-level intervals 8,000, and the simulation and the probability-based sensitivity analysis 1,000; the seed for each is fixed in the analysis scripts and listed in the repository. The R and Python implementations agree to machine precision on every deterministic quantity; the bootstrap intervals use separate random number streams and so agree only to Monte Carlo error, which at 2,000 replicates is about 8×10^{-5} on the interval bounds.

Simulation study

We simulated paired reader studies of 40,000 cases read by 40 clinicians. Each case carried a hidden severity, and the true state was drawn from a logistic model in that severity with intercept $\text{logit}(0.12)$ and slope 1.6, which gives a realised event rate of 0.197. The clinician's judgement and the AI's were drawn from the same severity with independent noise, with the AI the better of the two agents. The assisted judgement was a weighted blend of the two sets of log-odds, where the weight stands for how much the clinician defers. Clinicians acted when their probability passed a personal threshold near 0.10. The clinician's personal threshold was deliberately set far below the event rate, resulting in frequent treatment choices that lower baseline accuracy relative to a strategy of never acting.

Three scenarios were fixed in advance: a reassuring AI whose output is shifted downwards with heavy deference, a well calibrated AI with moderate deference, and an alarmist AI shifted upwards with heavy deference. Coverage of the intervals was checked in each scenario over 300 replicates of 1,500 cases against a value computed from 400,000 cases.

The Collab-CXR study

Collab-CXR is a public reader study in which 227 radiologists, with 22 years of experience on average, read 324 historical chest X-rays [22, 23]. For each of 104 findings they gave a probability and, where the finding allowed it, a binary treatment or follow-up recommendation, under four conditions crossing help from a CheXpert-trained model [24] with whether clinical history was shown. The reference standard is

the pooled judgement of five board certified radiologists at one US centre, and we used the binary label from their simple average. The data are released under CC-BY 4.0. The experiment was first reported by Agarwal and colleagues [25], and designs 1 and 2 were later analysed by Yu and colleagues [6].

All three designs are within radiologist. None of them randomises radiologists between arms. Design 1 has 159 radiologists recruited through teleradiology firms; it randomises the order of the four conditions but never shows the same image twice, so it gives Δ NB and not the split. Design 2 has 33 radiologists from a single Vietnamese health system, working without monetary incentive, each reading the same 60 cases under all four conditions in sessions at least two weeks apart; it is paired, counterbalanced and washed out. Design 3 has 35 radiologists from teleradiology firms, each reading the same 50 cases twice, first without the AI and then with it; it is paired, but the AI condition always came second and no washout is described. Clinical history was randomised to half of each radiologist's 50 cases, so holding history constant leaves 25 of them per radiologist in the analysis.

Design 2 is designated primary on design grounds — it is the only design that is paired, counterbalanced and washed out — a judgement that can be made without reference to any result. Designs 3 and 1 are reported as replications in a different population. We did not pool the three, because they differ in where the radiologists were recruited, whether they were paid, and how many times they saw each case.

One detail of the data structure has to be right or the analysis fails silently. The variable `uid_clean` labels a reading session, not a person: in design 2 each radiologist completes four sessions and gets a different value in each, so grouping on it makes design 2 look as though it contains no paired cells at all. The participant identifier is `experiment_id`, which is stable across sessions and gives 159, 33 and 35 participants in designs 1, 2 and 3, adding to the 227 radiologists the data descriptor reports. The reference-standard readings sit in the same file and were excluded.

The unit of analysis is one finding, in one patient, judged by one radiologist, because the recommendation in this study is recorded per finding rather than per patient. Net benefit is therefore per decision here, not per patient. We held clinical history constant at present and kept the findings that the model scores and that carry a recommendation. The model scores 15 of the 104 findings; support device hardware has no treatment question, which leaves 14, and two of those are fracture findings sharing one model output, so the 14 findings carry 13 distinct scores. Design 2 gives $33 \times 60 \times 14 = 27,720$ paired decisions at an event rate of 0.0547, and design 3 gives $35 \times 25 \times 14 = 12,250$ at an event rate of 0.0568.

We looked at thresholds from 0.005 to 0.25 and report at 0.02, 0.05, 0.10 and 0.20; the figures extend the grid to 0.2825 so that the break-even threshold falls inside the plotted range. That range was set by the clinical question, not by the results: the decision being recorded is whether to treat or follow up an abnormality on a chest film, where missing something matters far more than an unnecessary follow-up, so a clinician's threshold sits low.

Three of the 14 findings have no positive cases under the reference standard. We kept them, because a decision to act on them is still a real decision with a real cost, and dropping them would flatter the result. Sensitivity analyses without them, without the umbrella label, and using both clinical-history conditions are reported.

One assumption deserves to be said out loud. Pooling 14 findings under a single threshold treats the stakes as the same for all of them, and they are not. A radiologist will act on a suspected pneumothorax at a much lower risk than on atelectasis. The finding-level analyses avoid this, were not adjusted for multiple comparisons, and should be read as exploratory.

Use of a large language model

A large language model, Claude Opus 5 from Anthropic, was used to assist with mathematical derivations, writing and debugging the analysis software, and sentence refinement. The content was reviewed and edited by the authors; the authors take full responsibility for the content.

Results

Accuracy and clinical value can point opposite ways, in simulation

Table 1 and Fig. 1 give the simulation. In all three scenarios the unaided clinician's accuracy was 0.577 by construction, which is below the 0.803 you get by never acting at this event rate; that is a property of the generative model, not of clinician behaviour.

Table 1
Simulated paired reader studies.

Accuracy, unaided	A: reassuring	B: calibrated	C: alarmist
	0.577	0.577	0.577
Accuracy, assisted	0.738	0.603	0.508
Change in accuracy	+ 0.160	+ 0.025	-0.069
Decisions changed	25.2%	14.6%	20.2%
Moved to act, r_U	2.5%	6.8%	15.4%
Event rate there, π_U	0.249	0.186	0.135
Moved to stand down, r_L	22.7%	7.8%	4.8%
Event rate there, π_L	0.119	0.063	0.051
ΔNB at $p = 0.05$	-0.0113	+ 0.0087	+ 0.0138
95% CI	-0.0130, - 0.0097	+ 0.0073, + 0.0103	+ 0.0120, + 0.0160
$P(\Delta NB > 0)$	0.00	1.00	1.00
Number needed to consult	-88	115	72
40,000 cases, 40 readers, realised event rate 0.197. ΔNB and NNC at $p = 0.05$. P is the share of bootstrap replicates in which ΔNB was above zero.			

The reassuring AI gave the most impressive accuracy result in the whole study, a gain of 16 percentage points, and was the only one of the three that harmed patients. It got that accuracy by talking clinicians out of acting on nearly a quarter of all cases, and the condition was present in 11.9% of those, more than twice the 0.05 threshold the decision was being made at. By the break-even rule that is harmful, and heavily so. One net true positive was lost for every 88 patients shown to the AI.

The alarmist AI is the mirror image. Its accuracy was 6.9 points worse than the unaided clinician, which would get most tools rejected. But at a threshold of 0.05, where missing the condition is nineteen times worse than acting unnecessarily, over-calling was the right mistake to make, and it gave the largest clinical benefit of the three (Fig. 2). Neither conclusion can be reached from accuracy, sensitivity, specificity, or any measure of the model on its own. Where the reassuring AI's harm sits in its output range is shown in Fig. 3.

No single number summarises an assistance effect either. The reassuring AI harmed decisions below a threshold of 0.103 and helped above it, and the alarmist AI did the reverse, crossing zero at 0.174. Eq. (5) reproduces both crossings to the precision of the threshold grid. The curve is the result, not a point on it.

The estimators behaved. Cluster bootstrap intervals covered the true value 92.3% to 95.3% of the time at a threshold of 0.05, 93.3% to 94.7% at 0.10, and 90.7% to 94.0% at 0.20, against a nominal 95%, with the worst figure in the reassuring scenario at the highest threshold, where the two switch sets are most lopsided. The closed form in Eq. (4) reproduced the direct computation to machine precision across randomly drawn parameters.

Collab-CXR: judging by accuracy would have thrown out a tool that helps

Across 27,720 paired decisions by 33 radiologists on 14 findings, the AI made the team less accurate. Accuracy fell from 0.8465 to 0.8382, a loss of 0.83 percentage points, because a drop in specificity from 0.8544 to 0.8428 outweighed a gain in sensitivity from 0.7093 to 0.7601.

Net benefit moved the other way, at every threshold below 0.20 (Table 2, Fig. 4). At $p = 0.05$, ΔNB was + 0.0022 (95% CI + 0.0007 to + 0.0038). That is 0.22 extra net true positives per 100 decisions, or one for every 455 decisions shown to the AI. In absolute terms it is small. In relative terms it is not, because the quantity it is added to is also small: the unaided radiologist’s net benefit at that threshold was 0.0316, so assistance added 7% to the clinical value the radiologist was already producing. We make no claim that this is a large benefit. The claim is about the sign, and about the fact that the two measures disagree on it. A breakdown of this benefit across individual findings is shown in Fig. 5.

Table 2
Collab-CXR design 2, the primary analysis.

Net benefit, treat all	p = 0.02	p = 0.05	p = 0.10	p = 0.20
	0.0354	0.0050	-0.0503	-0.1816
Net benefit, unaided	0.0360	0.0316	0.0235	0.0044
Net benefit, assisted	0.0386	0.0338	0.0251	0.0044
ΔNB	+ 0.0026	+ 0.0022	+ 0.0016	+ 0.0000
95% CI	+ 0.0010, + 0.0042	+ 0.0007, + 0.0038	+ 0.0000, + 0.0031	-0.0019, + 0.0019
$P(\Delta NB > 0)$	1.00	1.00	0.98	0.52
Gain relative to unaided	+ 7%	+ 7%	+ 7%	0%
Number needed to consult	392	455	645	not reported
Change in accuracy	-0.0083	-0.0083	-0.0083	-0.0083
27,720 paired decisions by 33 radiologists, clinical history held constant. Intervals are 95% percentile bootstrap intervals clustered on the radiologist, from 2,000 replicates.				

The split into up-switches and down-switches explains the contradiction. The AI changed 2,851 of 27,720 decisions, which is 10.3%. It moved radiologists to act 1,617 times, and the finding was really there in 11.9% of those. It moved them to stand down 1,234 times, and the finding was really there in 9.4% of those. Because moving up happened more often and was right more often, the balance came out positive.

Accuracy cannot see this. It weighs a false positive and a false negative the same, so all it registers is that the radiologists said “treat” more often and were wrong more often in raw count. It records a loss. Net benefit knows that at a threshold of 0.05 a missed finding costs nineteen times what an unnecessary follow-up costs, so it prices the same switches differently and records a gain.

By Eq. (5) the effect crosses zero once, at $p^* = 0.201$. The point estimate stays positive up to p^* , but the interval clears zero only up to about 0.05, so the crossing marks where the sign of the estimate turns, not the limit of what the data support.

The intervals are wide enough

Clustering on the radiologist gave the widest interval at 0.02, 0.05 and 0.10; at 0.20 the patient-clustered interval was marginally wider (Table 3). We report the radiologist-clustered interval throughout. The conclusion does not depend on the choice: assistance is beneficial at 0.02 and 0.05 under all four schemes, at 0.10 every lower bound lies within 10^{-4} of zero, and at 0.20 the interval includes zero under all four.

Table 3
The net benefit of assistance under four resampling schemes, design 2.

p	Δ NB	By radiologist	By patient-finding	By patient	Unclassified
0.02	+ 0.00255	+ 0.00099, + 0.00416	+ 0.00128, + 0.00385	+ 0.00116, + 0.00404	+ 0.00130, + 0.00383
0.05	+ 0.00220	+ 0.00067, + 0.00377	+ 0.00089, + 0.00350	+ 0.00082, + 0.00370	+ 0.00096, + 0.00345
0.10	+ 0.00155	+ 0.00003, + 0.00314	+ 0.00017, + 0.00288	+ 0.00006, + 0.00313	+ 0.00024, + 0.00285
0.20	+ 0.00002	-0.00193, + 0.00189	-0.00163, + 0.00162	-0.00191, + 0.00196	-0.00153, + 0.00154
Each cell is the 95% percentile bootstrap interval under that scheme. Clustering on the radiologist gives the widest interval at 0.02, 0.05 and 0.10; at 0.20 the patient-clustered interval is marginally wider. The radiologist-clustered interval is what we report throughout.					

Thirty three radiologists is not many to resample, so we also ran a cluster sign-flip permutation test over 100,000 sign vectors. The observed Δ NB fell outside the central 95% of the permutation distribution at 0.02 (two-sided $p = 0.004$) and at 0.05 ($p = 0.011$), fell just inside it at 0.10 ($p = 0.071$), and sat near its centre at 0.20 ($p = 0.99$). The permutation check is therefore stricter than the percentile interval at 0.10,

where the margin from zero is in any case smaller than the Monte Carlo error between our two implementations, and we rest the claim on thresholds up to about 0.05.

The effect is not the work of a few readers

A switch rate of 10.3% could come from every radiologist changing a few decisions or from a handful changing many, and the two would mean different things. Across the 33 radiologists the median switch rate was 10.4%, with a middle half from 8.2% to 12.4% and a full range from 2.1% to 21.2%. Every radiologist changed at least one decision. The seven most influenced radiologists, the top fifth, accounted for 31% of all switches against the 20% you would expect if influence were uniform. Influence is uneven, as one would expect, but it is not the work of a small group.

Where the radiologist beats acting on everyone, and where they do not

An assistance curve compares a clinician with themselves, so it can be positive in a region where both arms are beaten by a blanket policy. That happens here at the bottom of the range, and it should be said plainly. At $p = 0.02$ the net benefit of acting on every finding is 0.0354, just below the unaided radiologist's 0.0360 and the assisted radiologist's 0.0386. Acting on everything beats the unaided radiologist below $p = 0.019$ and the assisted radiologist below $p = 0.016$, so below about 0.016 the comparison is between two strategies that are both worse than a blanket policy, and the positive ΔNB there is not a reason to deploy anything. That is a property of these radiologists on these cases at a very low threshold, not of the method, and assistance shrank the region by moving radiologists towards acting in exactly the range where acting more was correct. We rest the conclusion on $p = 0.02$ and above.

The result holds in two other groups of radiologists

Design 3 is 12,250 paired decisions by 35 radiologists recruited through teleradiology firms. Accuracy fell from 0.8946 to 0.8852, a loss of 0.94 percentage points, while ΔNB was + 0.0046 at $p = 0.02$, + 0.0041 at 0.05 (95% CI + 0.0007 to + 0.0077) and + 0.0033 at 0.10. The AI changed 959 decisions, 7.8%, moving radiologists to act 597 times with the finding really there in 18.4% of those, and to stand down 362 times with it really there in 13.8%. The break-even threshold was 0.255.

Design 1 is 66,780 decisions by 159 radiologists. No image is read twice, so only ΔNB is available. Accuracy fell from 0.8884 to 0.8818 and ΔNB was + 0.0047 at $p = 0.05$ (95% CI - 0.0000 to + 0.0093). Because no image is read twice, the two arms contain different images, and they are not balanced: the assisted arm has an event rate of 0.0529 against 0.0497 unassisted, 1,767 positive cases against 1,658. Net benefit rises with the event rate whatever the clinician does, and the difference in treat-all net benefit between the arms is + 0.0034 at $p = 0.05$, about three quarters of the estimate. Standardising both arms to the pooled event rate using each arm's own sensitivity and specificity gives + 0.0027. The design 1 figure should be read with that in mind.

All three designs point the same way (Fig. 6). Both paired designs exclude zero at $p = 0.05$ and the unpaired one does not, even though design 1 has more than twice the decisions and nearly five times the

radiologists of design 2. Pairing removes the variance between readers, and that is worth more here than the extra sample size. The primary design also gave the smallest of the three estimates, and each replication has a reason to be larger. In design 3 the AI condition always came second, so any order effect is aligned with it, although the original investigators tested for order effects on diagnostic quality and found none. In design 1 the arms are unbalanced, and the imbalance alone accounts for most of the gap. The design that controls for both gives the most conservative answer.

Sensitivity analyses

Dropping the three findings with no positive cases leaves 21,780 decisions across 11 findings at an event rate of 0.0697, and the gap widens on both sides: accuracy falls by 1.13 points rather than 0.83, ΔNB rises to + 0.0028 at $p = 0.05$ (95% CI + 0.0008 to + 0.0048), the number needed to consult improves to 363, and the break-even threshold moves to 0.192. Keeping the empty findings is therefore the conservative choice, which is why it is the main analysis.

Dropping the umbrella label, which contains the other thirteen findings and so enters the pooled estimate twice, leaves 25,740 decisions and ΔNB of + 0.0020 (95% CI + 0.0007 to + 0.0035). Using both clinical-history conditions gives 55,440 decisions and ΔNB of + 0.0023 (95% CI + 0.0011 to + 0.0035).

Repeating the analysis on the radiologists' own probability judgements, with the action defined as a probability above the threshold, holds the action rule fixed at p and so avoids the assumption above. It gave the same direction and a smaller estimate, + 0.0011 at $p = 0.05$ (95% CI - 0.0000 to + 0.0023). The recorded recommendation is the clinically consequential act and remains our primary outcome, but the two should be read together.

The answer differs by finding, and one number hides it

The pooled figure averages over 14 findings that behave differently (Fig. 5). Among the 11 of the 14 findings that have at least one positive and one negative case under the reference standard, each contributing 1,980 decisions, seven showed the accuracy change and the clinical benefit pointing in opposite directions, and three did so by a margin we set in advance as material, meaning an accuracy change of at least 0.005 together with an absolute ΔNB of at least 0.001: the umbrella label, airspace opacity and atelectasis. Those cut-offs are arbitrary and the count is sensitive to them.

Three findings had intervals excluding zero. Two are clear by any reading, and both are cases where the AI moved radiologists towards acting on genuinely abnormal images; the third, pleural effusion at + 0.0039 (95% CI + 0.0001 to + 0.0078), has a lower bound within 10^{-4} of zero and we rest nothing on it. For atelectasis, ΔNB was + 0.0131 (95% CI + 0.0040 to + 0.0222) while accuracy fell by 0.66 points; the AI moved radiologists to act 152 times and the finding was really there 44 times, an event rate of 0.29 in the up-switched cases against a background rate of 0.12. For airspace opacity, ΔNB was + 0.0113 (95% CI + 0.0036 to + 0.0196) while accuracy fell by 3.0 points.

Nothing was clearly harmful. Cardiomedial abnormality had the most negative estimate, -0.0024 (95% CI -0.0082 to $+0.0034$), and oedema was -0.0016 (95% CI -0.0044 to $+0.0012$). Oedema is worth a sentence because it was the sharpest result in design 3, where ΔNB was -0.0037 (95% CI -0.0070 to -0.0010) and not one of the 74 cases in which the AI persuaded a radiologist to act on oedema was a true positive. In the primary analysis 4 of 215 such cases were true positives, so that did not replicate, and we report it as a design 3 observation rather than a property of the model.

The override profile found little, and that is worth reporting

At $p = 0.05$ the profile found four bands of model output with a negative contribution, all of them small. Suppressing them would move ΔNB from $+0.0022$ to $+0.0024$, a gain of about 9%. At $p = 0.02$ only one band is negative and suppression gains 3.5%. Unlike the simulated example (Fig. 3), no display rule here changes the verdict, because this model's harm is not concentrated in one part of its output range. We report the null on purpose. A method that always finds something is a method to distrust.

Discussion

We set out to measure something the field has been describing for five years without being able to quantify. The result on real radiologists was not the one we expected.

In the most rigorous of the three designs, a CheXpert-trained model made 33 radiologists less accurate and their decisions better, at the same time, on the same decisions. The direction held in two more designs covering 194 further radiologists, and survived four ways of resampling the data, a permutation check and four sensitivity analyses. Judged the way clinical AI is normally judged, this tool made experienced doctors worse. Judged by what happened to the patients in front of them, it helped, clearly so for any decision a clinician would act on at a risk between roughly one in fifty and one in twenty, and in the same direction though less certainly above that. Neither number is wrong about what it measures. Only one of them is measuring the thing that happens to patients.

This reconciles a literature that looks contradictory

Reader studies of clinical AI reach conclusions that are hard to line up. Some report benefit on average, though the clearest of those also found accuracy degraded on the cases where the model erred [26]. Some report benefit only when the model is correct and harm when it is not [1, 2, 4, 5]. Some report no change in the average with large offsetting effects underneath [5, 6]. If a study reports accuracy, it is reporting a quantity whose sign can differ from the clinical effect on the same data. Two honest studies of the same tool, one where misses are very costly and one where over-treatment dominates, will disagree, and nothing in the usual reporting makes that visible. Assistance curve analysis does not settle those disagreements by declaring a winner. It makes them visible, by forcing the price of the two errors into the open and showing where along the range of thresholds the sign changes.

The value of an AI lives on the decisions it changes

Result 1 makes this exact, and it has an uncomfortable consequence for how these systems are built. Effort spent improving a model's performance on cases where the clinician was always going to be right cannot change any patient's outcome. In our primary analysis, 89.7% of decisions were the same with and without the AI and contributed exactly nothing.

The question during development is therefore not how accurate the model is. It is which decisions it will be believed on, and whether it is right there. Work on selective deferral points the same way. Dvijotham and colleagues showed that a system which learns when to defer to the clinician can beat both the model alone and the clinician alone where the two are complementary, and that where they are not, as in two of five commercial tuberculosis systems they tested, no confidence-based deferral rule would have helped [27]. Our override profile is the matching idea on the evaluation side, asking after the fact which part of a model's output range earned its place on the screen.

One number for a device is the wrong object

The finding-level results are the half of this paper that worries us most. The same model, shown to the same radiologists, reading the same images, ranged from clearly good for atelectasis to slightly harmful on the point estimate for cardiomeastinal abnormality.

Regulators approve devices. Hospitals buy devices. Journals report accuracy for devices. But the unit that matters clinically is a decision about a particular question, and the sign of the effect is not constant across those questions. Anything reported at the level of the device is averaging over a quantity that changes sign, which is precisely the failure Kiani and Yu described [5, 6] and precisely what a single accuracy figure cannot show.

Reporting guidelines are the natural place to fix this, and less needs adding than one might think. DECIDE-AI item 12 already asks for "any instances of and reasons for user variation from the AI system's recommendations and, if applicable, users changing their mind based on the AI system's recommendations" [28]. Those are exactly the cases where clinical value is made. What the checklist does not yet ask for is what they were worth: a decision-analytic measure of the assisted decision, reported across the plausible range of thresholds and broken down by clinical question. That is an amendment to an item that already exists.

Limitations

Our method measures the decision, not what happened to the patient afterwards, and it rests on the standard assumption of decision analysis, that the threshold probability captures the trade-off honestly. It also says nothing about the slower effects: Budzyń and colleagues found unassisted detection rates lower in the three months after AI arrived [8], and a reader study measures the decision made with the tool on the day, not what the tool is doing to the clinician over a year.

The method needs a known truth for each case, so it works only where a reference standard exists. It assumes a single binary action; an extension of decision curve analysis to several treatment options

exists [15], but it uses a different net benefit functional based on risk differences and needs trial or network meta-analytic data, so it is not a drop-in generalisation. The split into up-switches and down-switches requires a paired design, which is a real constraint on study planning, though it buys the ability to explain a result rather than only report it.

The applied analysis has limits of its own. It is one study, one model, one imaging test and one reference standard, and that standard is a panel of five radiologists rather than tissue or an outcome, so what we measure is agreement with expert consensus. If the panel and the model share a blind spot, this analysis cannot see it. The model was trained on CheXpert labels, which are derived from radiology reports by an automated labeller, and the reference standard is expert consensus on the same fourteen-observation taxonomy. Agreement with the panel therefore partly reflects a shared convention about what each finding means, and any systematic disagreement between that convention and the underlying pathology is invisible to this analysis. The reading was done on historical cases outside a working clinic. The primary analysis rests on 33 radiologists from a single health system with no incentive for accuracy, and the cluster bootstrap resamples those 33 units, which is the effective sample size.

The unit of analysis is a finding rather than a patient, and one patient contributes 14 correlated decisions, one of which contains the other thirteen. The patient-clustered bootstrap addresses that dependence and gives a narrower interval than the one we report at every threshold on which the conclusion rests, and the analysis without the umbrella label leaves the conclusion unchanged. Acting on every finding of every patient is not a policy anyone would adopt, so the treat-all reference is less meaningful here than at the patient level. Our finding-level analyses have 1,980 decisions each, the intervals are correspondingly wide, and no correction was made for multiple comparisons, so the count of three material sign disagreements shows that disagreement is common rather than estimating how common.

One reanalysis cannot say how often accuracy and clinical value part company across the literature. Answering that needs a systematic reanalysis of published reader studies, and the data it requires, which is the truth together with the unaided and the aided decision for every case, is rarely reported so far as we have been able to establish. Collab-CXR is the exception, and it is the only reason the applied half of this paper exists. We would ask anyone running a reader study to deposit those three columns, and to record a participant identifier that stays the same across sessions.

Conclusions

An AI that never changes a clinician's mind has no clinical value however accurate it is. An AI that changes minds in the wrong direction has negative value however accurate it is. Neither fact is visible in the measures the field currently reports.

Assistance curve analysis measures what is created on the decisions that move, prices the two errors the way the clinical question prices them, and returns a verdict together with the range of thresholds over which the aid is worth having. It needs nothing beyond what a paired reader study already collects. In the

one large public study where those records exist, it found that the standard way of judging a decision aid would have rejected a tool that helps patients, and would have hidden the fact that the size of the help, and sometimes its sign, changes from one finding to the next.

Abbreviations

AI
artificial intelligence
CI
confidence interval
CXR
chest radiograph
FP
false positive
NB
net benefit
 Δ NB
net benefit of assistance
NNC
number needed to consult
OR
odds ratio
TP
true positive.

Declarations

Ethics approval and consent to participate.

Not required. This is a methods study and a secondary analysis of a publicly available, de-identified dataset released under CC-BY 4.0.

Consent for publication.

Not applicable.

Competing interests.

The authors declare that they have no competing interests.

AI declaration.

During the preparation of this work, the authors used the Opus 5 model from Anthropic to assist with mathematical derivation, writing and debugging of the analysis software, and sentence refinement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding.

This work received no funding.

Author Contribution

MA and YMN conceived the method, derived the theory, wrote the software with assistance of AI, carried out the simulation study and the applied analysis, and wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgement

We thank the Collab-CXR investigators for releasing decision-level data, which is what made the applied part of this work possible.

Data Availability

The dataset supporting the conclusions of this article is available in the Open Science Framework repository under CC-BY 4.0, <https://doi.org/10.17605/OSF.IO/Z7APQ> [22,23]. The assistcurve package for R and Python, the simulation study, the permutation check, the derived analysis files and all code needed to reproduce every figure, table and reported number are available as follows: Project name: assistcurve Project home page: <https://github.com/Marwan-Alqumbaey/assistcurve> Archived version: v0.1.0, <https://doi.org/10.5281/zenodo.22240681> Operating system(s): platform independent Programming languages: Python and R Other requirements: Python 3.9 or higher with numpy; pandas and matplotlib are additionally required for the analysis scripts and figures. R 3.6 or higher, base R only to use the package; testthat to run its test suite Licence: MIT Any restrictions to use by non-academics: none

References

1. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. *Nat Med.* 2020;26(8):1229–34. 10.1038/s41591-020-0942-0.

2. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lerner E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med.* 2021;4:31. 10.1038/s41746-021-00385-9.
3. Kücking F, Busch DA, Przysucha M, Kutza J-O, Hannemann N, Hüsters J, et al. Impact of AI recommendation correctness on diagnostic accuracy in clinical decision-making. *Int J Med Inf.* 2026;207:106223. 10.1016/j.ijmedinf.2025.106223.
4. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *JAMA.* 2023;330(23):2275–84. 10.1001/jama.2023.22295.
5. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit Med.* 2020;3:23. 10.1038/s41746-020-0232-8.
6. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat Med.* 2024;30(3):837–49. 10.1038/s41591-024-02850-w.
7. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology.* 2023;307(4):e222176. 10.1148/radiol.222176.
8. Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol.* 2025;10(10):896–903. 10.1016/S2468-1253(25)00133-5.
9. Van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *Lancet Digit Health.* 2025;7(12):100916. 10.1016/j.landig.2025.100916.
10. Schemmer M, Kühl N, Benz C, Bartos A, Satzger G. Appropriate reliance on AI advice: conceptualization and the effect of explanations. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces.* New York: ACM; 2023. pp. 410 – 22. 10.1145/3581641.3584066
11. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med.* 1975;293(5):229–34. 10.1056/NEJM197507312930505.
12. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med.* 1980;302(20):1109–17. 10.1056/NEJM198005153022003.
13. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak.* 2006;26(6):565–74. 10.1177/0272989X06295361.
14. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res.* 2019;3:18. 10.1186/s41512-019-0064-7.
15. Chalkou K, Vickers AJ, Pellegrini F, Manca A, Salanti G. Decision curve analysis for personalized treatment choice between multiple options. *Med Decis Mak.* 2023;43(3):337–49. 10.1177/0272989X221143058.

16. Netto Flores Cruz G, Korthauer K. Bayesian decision curve analysis with bayesDCA. *Stat Med*. 2024;43(30):6042–58. 10.1002/sim.10277.
17. Vickers AJ, Van Calster B, Wynants L, Steyerberg EW. Decision curve analysis: confidence intervals and hypothesis testing for net benefit. *Diagn Progn Res*. 2023;7:11. 10.1186/s41512-023-00148-y.
18. Vickers AJ, Van Calster B, Wynants L, Steyerberg EW. Correction: decision curve analysis: confidence intervals and hypothesis testing for net benefit. *Diagn Progn Res*. 2025;9:8. 10.1186/s41512-025-00188-6.
19. Sunwoo L, Ryu WS, Buch K, Conklin J, Mehan W, Desalvo MN, et al. AI assisted detection of large vessel occlusion on non-contrast CT: multinational validation and reader study. *J Neurointerv Surg*. 2026. 10.1136/jnis-2026-025339.
20. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74(6):796–804. 10.1016/j.eururo.2018.08.038.
21. Baker SG. The summary test tradeoff: a new measure of the value of an additional risk prediction marker. *Stat Med*. 2017;36(28):4491–4. 10.1002/sim.7132.
22. Moehring A, Kutwal M, Huang R, Banerjee O, Jacobi A, Eber C, et al. A dataset for understanding radiologist-artificial intelligence collaboration. *Sci Data*. 2025;12:739. 10.1038/s41597-025-05054-0.
23. Moehring A, Collab. Open Sci Framew. 2025. <https://doi.org/10.17605/OSF.IO/Z7APQ>. -CXR: a dataset for understanding radiologist – artificial intelligence collaboration.
24. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Illcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc AAAI Conf Artif Intell*. 2019;33(1):590–7. 10.1609/aaai.v33i01.3301590.
25. Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining human expertise with artificial intelligence: experimental evidence from radiology. Cambridge (MA): National Bureau of Economic Research; 2023. Working Paper 31422. 10.3386/w31422
26. Reverberi C, Rigon T, Solari A, Hassan C, Cherubini P, GI Genius CADx Study Group. Experimental evidence of effective human-AI collaboration in medical decision-making. *Sci Rep*. 2022;12:14952. 10.1038/s41598-022-18751-2.
27. Dvijotham K, Winkens J, Barsbey M, Ghaisas S, Stanforth R, Pawlowski N, et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat Med*. 2023;29(7):1814–20. 10.1038/s41591-023-02437-x.
28. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924–33. 10.1038/s41591-022-01772-9.

Figures

Accuracy and clinical value disagree, in both directions

Top: decision curves. Bottom: net benefit of assistance with 95% cluster-bootstrap interval; red shading marks thresholds at which the AI makes decisions worse.

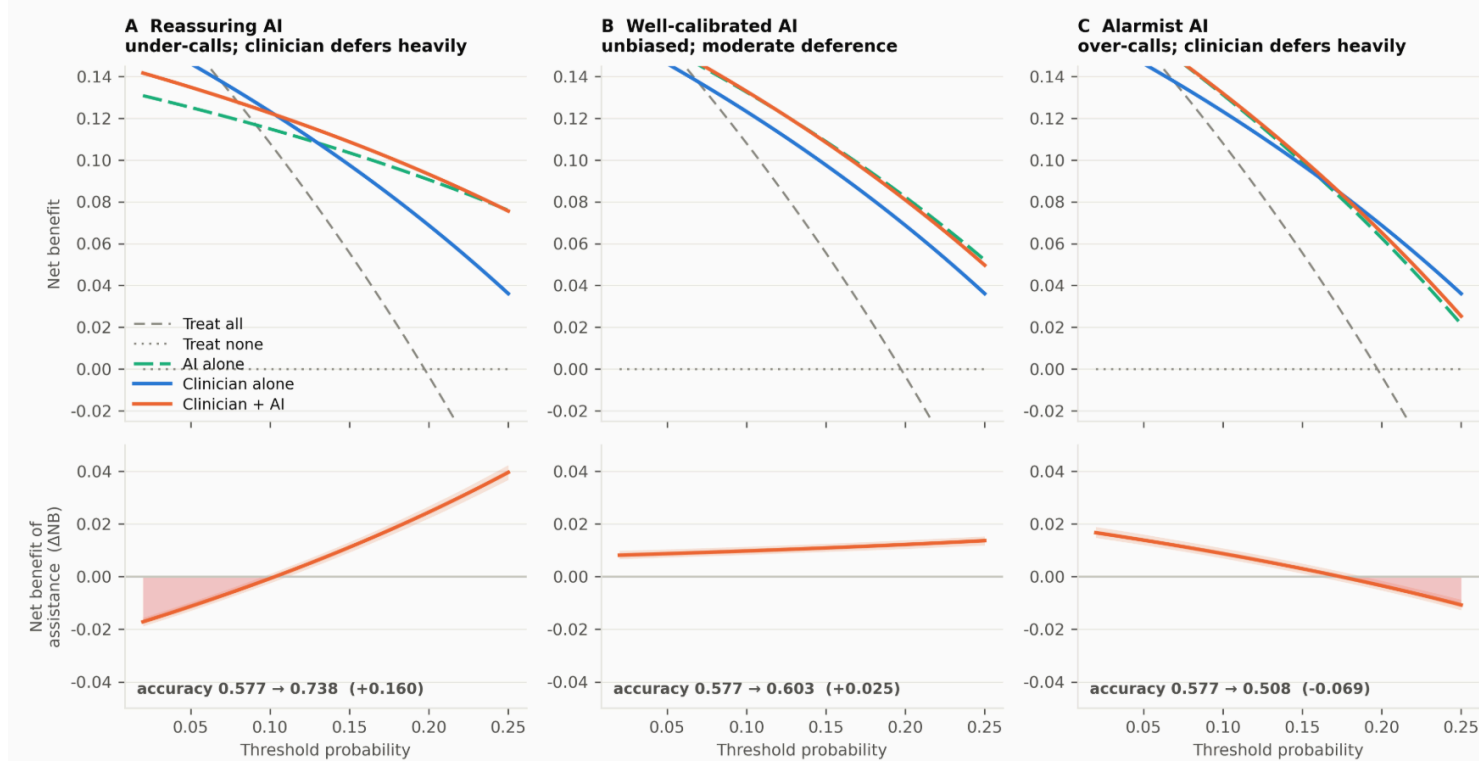


Figure 1

Assistance curves for three simulated AI systems. Top row: ordinary decision curves for treat-all, treat-none, the AI alone, the unaided clinician, and the clinician with the AI. Bottom row: the net benefit of assistance with its 95% cluster bootstrap interval, with red shading where the AI made decisions worse. The change in team accuracy is written on each panel. In panels A and C, accuracy and clinical value disagree in sign, in opposite directions.

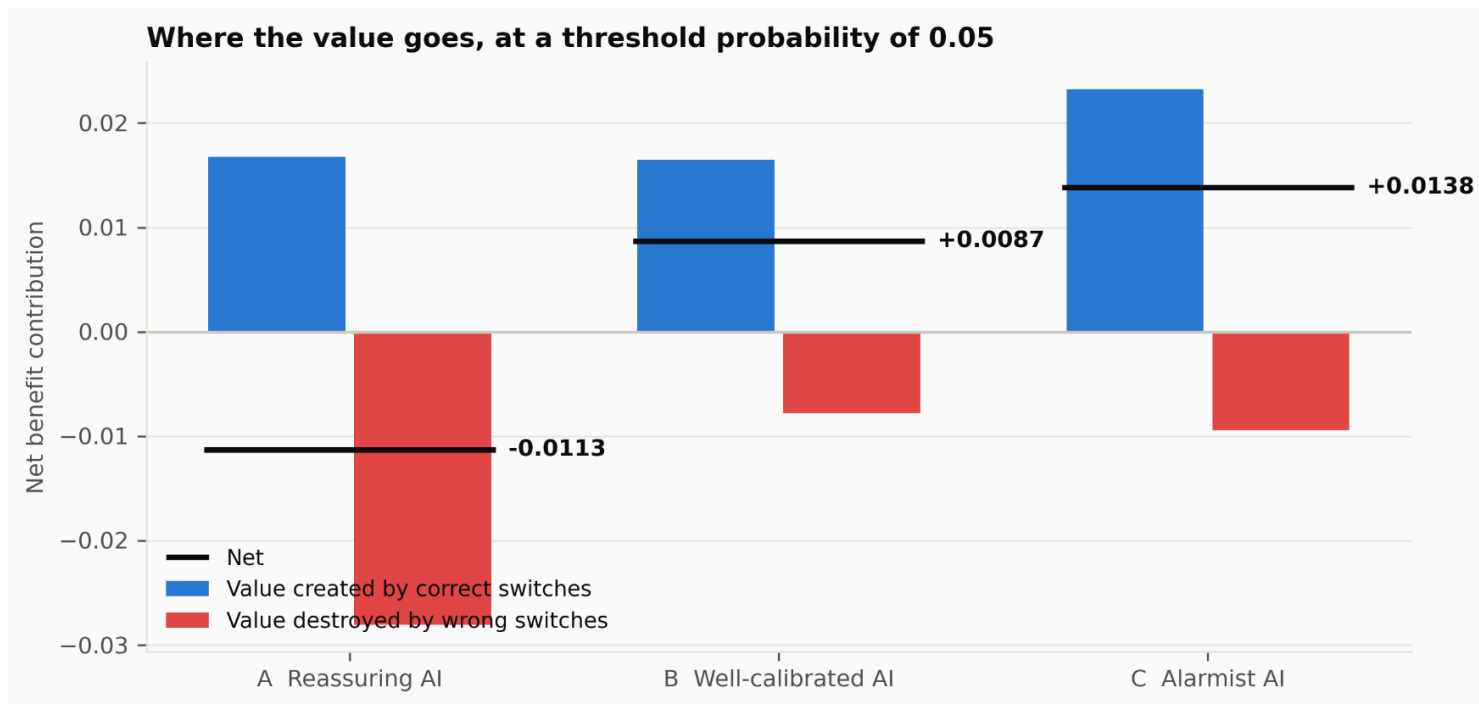


Figure 2

Where the value goes, at a threshold of 0.05. Blue is net benefit created by decision changes that were right. Red is net benefit destroyed by changes that were wrong. The black rule is the balance.

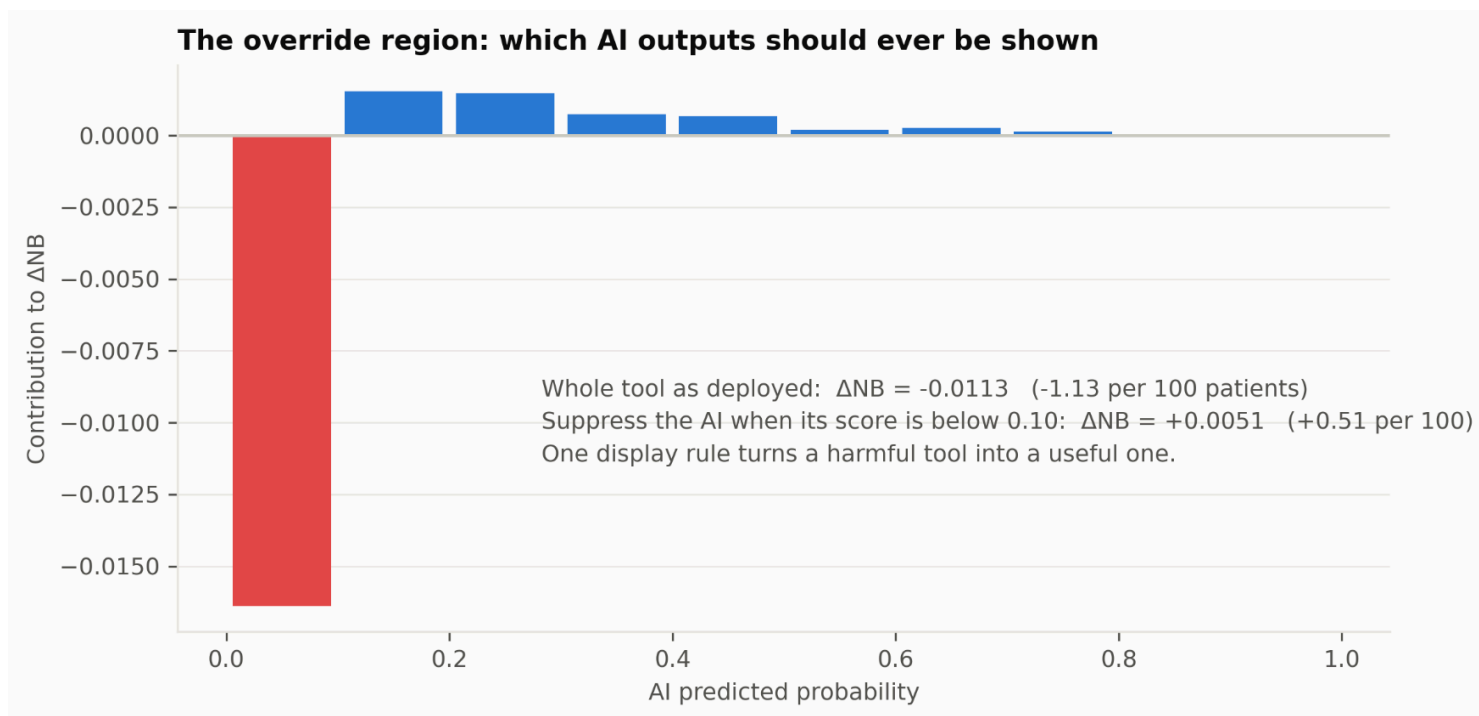


Figure 3

The override region for a reassuring simulated AI. Each band of the AI's predicted probability and what it contributed to ΔNB , at a threshold of 0.05. All the harm sits in the lowest band; suppressing the AI below

0.10 moves ΔNB from -0.0113 to $+0.0051$ without changing the model.

CheXpert made radiologists less accurate and their decisions better

Collab-CXR design 2, primary analysis: 27,720 paired decisions by 33 radiologists. Shaded band is the 95% cluster bootstrap interval over radiologists.

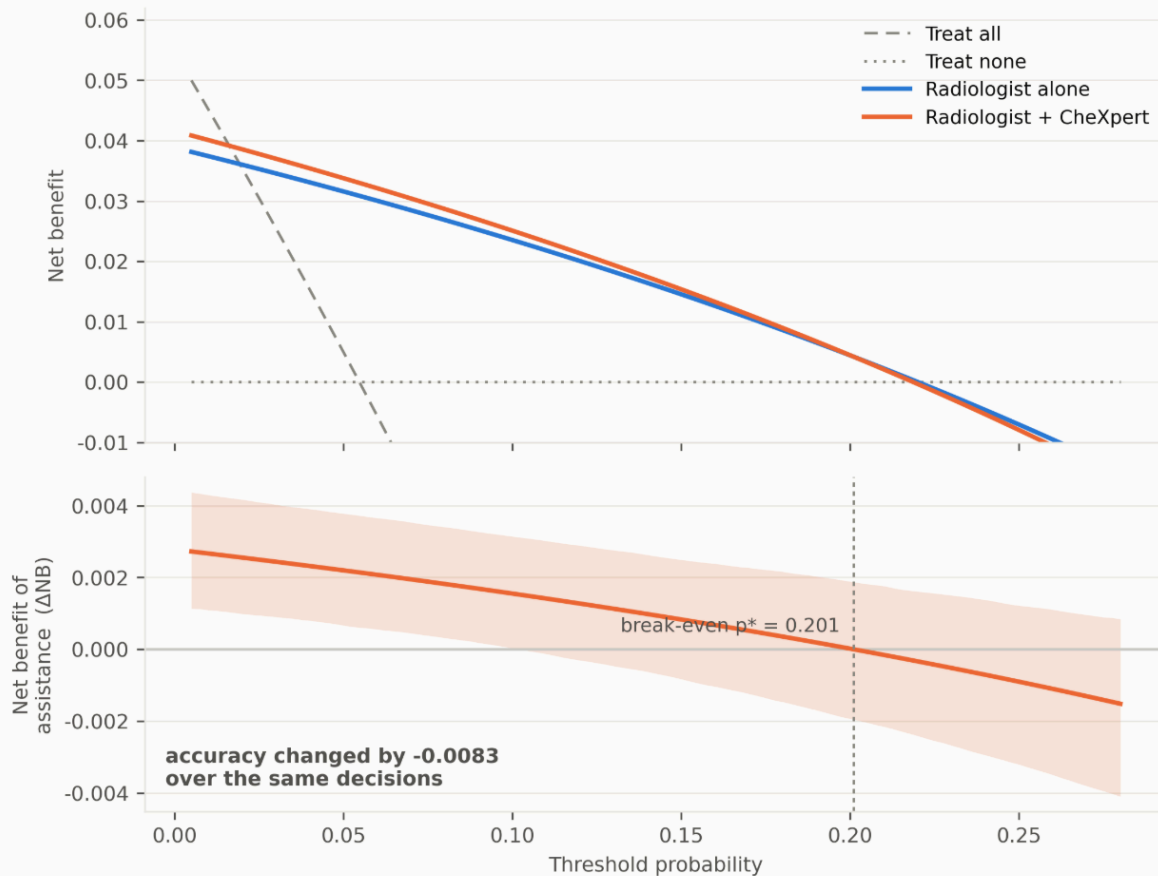


Figure 4

Assistance curve for Collab-CXR design 2, the primary analysis. 27,720 paired decisions by 33 radiologists with clinical history held constant, in the only design that is paired, counterbalanced and separated by a two-week washout. Top: net benefit for treat-all, treat-none, the unaided radiologist and the assisted radiologist. Bottom: the net benefit of assistance with its 95% cluster bootstrap interval over radiologists and the break-even threshold at 0.201. Over the same decisions, accuracy fell by 0.83 percentage points.

The two columns disagree for 3 of 11 findings

Same radiologists, same images, same model. Bars on the left are 95% cluster bootstrap intervals.

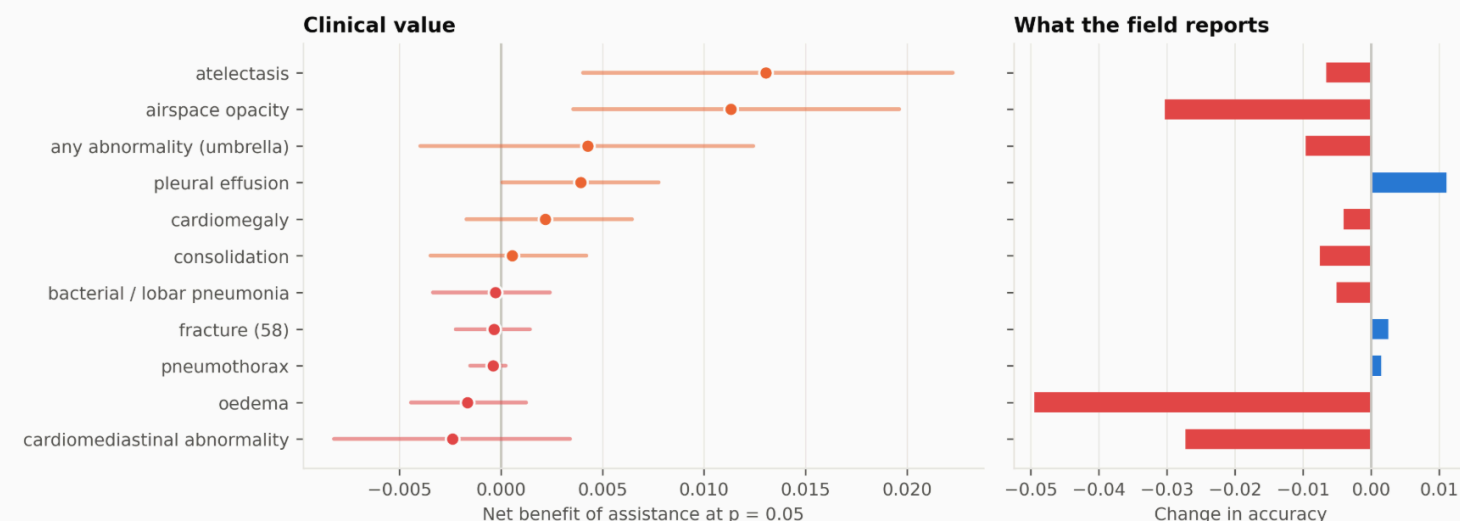


Figure 5

Net benefit of assistance by finding, primary analysis. Results at $p = 0.05$ for each finding in the primary analysis, with 95% cluster bootstrap intervals from 8,000 replicates, beside the change in accuracy for the same finding. Findings are ordered by ΔNB .

The effect holds in every design, and the most rigorous design is the most conservative

Design 2 is paired, counterbalanced and separated by a two-week washout. Design 3 is paired but the AI condition always came second. Design 1 reads each image once, so ΔNB is estimable but the switch decomposition is not. Bands are 95% cluster bootstrap intervals over radiologists.

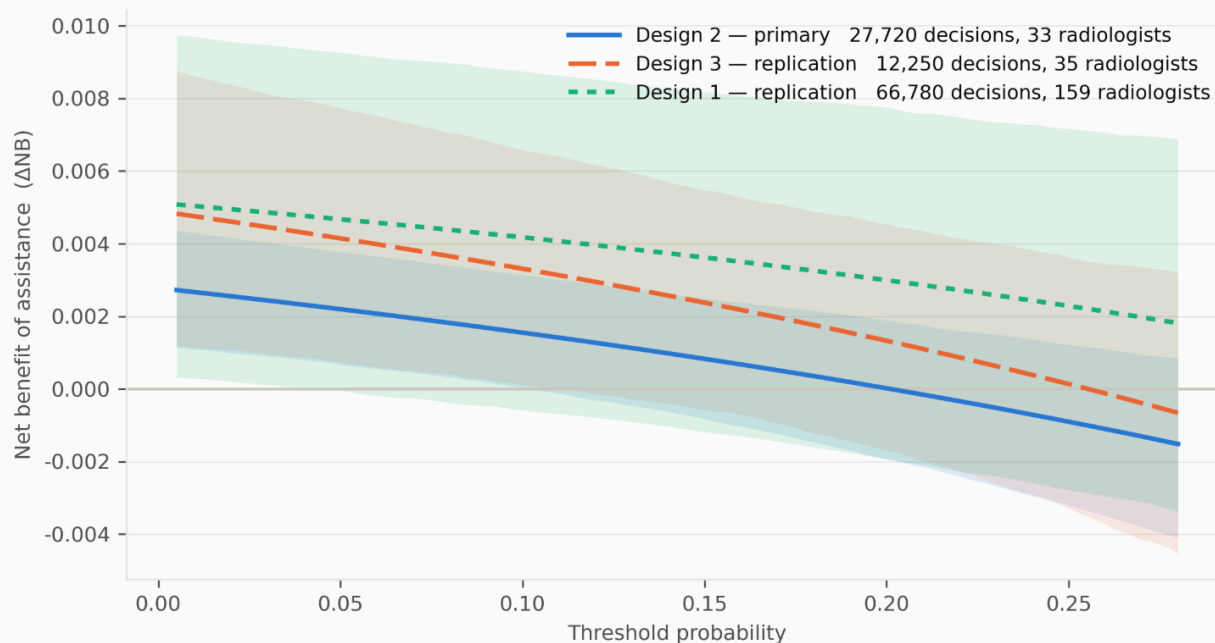


Figure 6

The net benefit of assistance in all three Collab-CXR designs. Bands are 95% cluster bootstrap intervals over radiologists. Design 2, the primary analysis, is paired, counterbalanced and washed out. Design 3 is paired but the AI condition always came second. Design 1 shows each image once, so ΔNB is available but the split into up-switches and down-switches is not.