


# Economics of AI and human task sharing for decision making in screening mammography

Received: 2 February 2024

Accepted: 12 February 2025

Published online: 07 March 2025

 Check for updatesMehmet Eren Ahsen <sup>1,2,6</sup> ✉, Mehmet U. S. Ayvaci <sup>3,6</sup>, Radha Mookerjee<sup>3</sup> & Gustavo Stolovitzky <sup>4,5</sup>

The rising global incidence of breast cancer and the persistent shortage of specialized radiologists have heightened the demand for innovative solutions in mammography screening. Artificial intelligence (AI) has emerged as a promising tool to bridge this demand-supply gap, with potential applications ranging from full automation to integrated AI-human decision-making. This study evaluates the economic feasibility of incorporating artificial intelligence (AI) into mammography screening within healthcare settings, considering full or partial integration. To evaluate the economic viability, we employ an optimization model specifically designed to minimize mammography screening costs. This model considers three distinct approaches when interpreting mammograms: automation strategy utilizing AI exclusively, delegation strategy involving the selective allocation of tasks between radiologists and AI, and the expert-alone strategy relying solely on radiologist decisions. Our findings underscore the significance of disease prevalence in relation to the trade-off between costs associated with false positives (e.g., follow-up expenses) and false negatives (e.g., litigation costs stemming from missed diagnoses) in shaping the AI strategy for healthcare organizations. We backtest our approach using data from an AI contest in which participants aimed to match or surpass radiologists' performance in assessing screening mammograms for women. The contest data supports the optimality of the delegation strategy, potentially leading to cost savings of 17.5% to 30.1% compared to relying solely on human experts. Our research provides guidance for healthcare organizations considering AI integration in mammography screening, with broader implications for work design and human-AI hybrid solutions in various fields.

Mammography plays a pivotal role in the early detection of breast cancer. In 2021, nearly 39 million women in the United States underwent mammography screening, highlighting the extensive scale and significant costs associated with this healthcare process. The growing global incidence of breast cancer and the limited availability of

specialized radiologists have created a demand-supply gap that AI promises to bridge<sup>1</sup>.

Can AI fully replace radiologists in interpreting mammograms for breast cancer screening? While AI has made remarkable advancements, recent experimental studies indicate that AI algorithms still

<sup>1</sup>Department of Business Administration, University of Illinois at Urbana-Champaign, Champaign, USA. <sup>2</sup>Department of Biomedical and Translational Sciences, University of Illinois at Urbana-Champaign, Champaign, USA. <sup>3</sup>Jindal School of Management, University of Texas at Dallas, Richardson, USA. <sup>4</sup>Department of Pathology, NYU Grossman School of Medicine, New York, New York, USA. <sup>5</sup>Biomedical Data Sciences Hub, NYU Langone Health, New York, New York, USA.

<sup>6</sup>These authors contributed equally: Mehmet Eren Ahsen, Mehmet U. S. Ayvaci. ✉e-mail: [ahsen@illinois.edu](mailto:ahsen@illinois.edu)

underperform compared to radiologists in mammography-based cancer screening<sup>2</sup>. While AI shows promise for automating mammography-based breast cancer screening, further research is required before it can be fully integrated into clinical practice<sup>1</sup>. In addition, rather than entirely replacing radiologists, AI can serve as a powerful tool for triaging and prioritizing cases. AI tools, such as advanced machine learning algorithms, can effectively identify mammograms with no significant findings, streamlining the review process, reducing radiologists' workload, and generating cost savings. For mammograms flagged with potential abnormalities, AI can promptly refer cases to radiologists for further evaluation, ensuring timely and accurate patient care. Supporting this approach, a recent Swiss clinical trial demonstrated that AI-assisted screening achieved cancer detection rates comparable to double-reading by radiologists, highlighting AI's potential as a triaging tool for human evaluation<sup>3</sup>. However, the economic implications of such a strategy remain unexplored.

In this research, we evaluate the economic viability of integrating AI-driven solutions into breast cancer screening programs and assess the cost and performance outcomes against the status quo, radiologist-only solutions. We examine the circumstances in which AI could completely replace radiologists and explore a strategy in which an AI algorithm provides an initial assessment, delegating specific cases to radiologists. To achieve our research objective, we formulate and solve an optimization model comparing three strategies: the current practice of relying solely on expert radiologists (expert-alone), the automation strategy where AI completely substitutes for radiologists, and the delegation strategy where AI and radiologists share responsibilities, with AI assigning specific cases to radiologists. Our model is grounded in the statistical principles of predictive AI and the economic implications of prediction-based decisions. To empirically validate our results, we leverage real-world mammography data from a mammography crowdsourcing challenge. The aim is to demonstrate how a healthcare organization can design its mammography operations, allocate tasks between radiologists and AI algorithms, and quantify the outcomes of such an arrangement.

## Results

### Analytical Model

We evaluate three distinct strategies for decision-making in mammography, each incorporating different degrees of AI involvement. The first strategy, reflecting current hospital practice, is the expert-alone approach, where radiologists independently classify patients as either sick (*s*) or healthy (*h*) based on mammograms. The second strategy, delegation, employs an initial AI screening to identify cases, with certain instances delegated to radiologists for final evaluation. In the delegation strategy, the AI algorithm evaluates the risk of breast cancer using mammography data and generates a continuous risk score, *r*, as a numerical value. If the estimated risk *r* is lower than a defined threshold *t<sub>D</sub>*, the algorithm classifies the patient as healthy (*h*), concluding the screening process. If the risk score *r* exceeds the threshold *t<sub>D</sub>*, the case is referred to a human expert for further evaluation and the final classification as either sick (*s*) or healthy (*h*). The rationale for this threshold-based approach is that very high-risk cases are more likely to indicate cancer, necessitating human intervention. In other words, the AI handles only the very low-risk cases. Delegating high-risk cases solely to AI would undermine the principle that human expertise is essential for follow-up procedures. Let *η* represent the assessment made by the human expert. Then, the decision rule for the delegation strategy, *d(r)* is given below:

$$d(r) = \begin{cases} h & \text{if } r < t_D \\ \eta \in \{h, s\} & \text{otherwise.} \end{cases} \quad (1)$$

Observe that when the threshold *t<sub>D</sub>* approaches to  $-\infty$ , the delegation strategy converges to the human-only strategy, as every case is

directed to the human expert for assessment, independent of the AI-generated risk score.

The third strategy, referred to as the automation strategy, eliminates the need for a human expert by relying entirely on the AI algorithm. The AI evaluates each mammogram and generates a risk score, *r*. When the generated score *r* is less than or equal to a predetermined threshold *t<sub>A</sub>*, the patient is labeled as healthy (*h*). If the risk score exceeds the threshold *t<sub>A</sub>*, the patient is labeled as sick (*s*). The automation strategy's decision rule *a(r)* is given below:

$$a(r) = \begin{cases} h & \text{if } r \leq t_A \\ s & \text{otherwise.} \end{cases} \quad (2)$$

Regardless of the chosen strategy—expert alone, delegation, or automation—no additional follow-up is required for patients classified as healthy (*h*). However, when a patient is classified as sick (*s*), follow-up procedures (additional imaging or biopsy), may be conducted. The disease prevalence in the population is represented by *λ*.

The classification outcome falls into one of four categories: true positive (*TP*), false positive (*FP*), true negative (*TN*), or false negative (*FN*). Let *P(o)* represent the probability of a specific outcome, where *o* ∈ {*TP*, *TN*, *FP*, *FN*}. These probabilities collectively determine the overall performance of the breast cancer screening system, whether it is based on the expert alone, delegation, or automation strategy. An effective screening system aims to maximize *P(TP)* and *P(TN)* while minimizing *P(FP)* and *P(FN)*.

Since radiologists do not provide exact risk assessment values prior to making their final decisions, we directly quantify the probabilities associated with each of the four possible outcomes, *P<sub>E</sub>(o)*, where *o* ∈ {*TP*, *TN*, *FP*, *FN*}. These probabilities are then used to assess the performance of the radiologists. These outcome probabilities represent the average performance of radiologists, capturing either the typical performance across a mix of radiologist expertise within a healthcare organization or the performance of an individual radiologist if a more personalized approach is adopted. In contrast, for algorithms, performance is assessed based on their generated risk assessment scores. We use the area under the ROC curve (AUC) to quantify the discriminatory power of an algorithm, which reflects its ability to effectively distinguish healthy patients from sick ones. In accordance with conventional approaches to statistical estimation and performance assessment in binary classification, we assume that the AI algorithm assigns risk scores that follow two distinct normal distributions: one for the healthy population and another for the sick population. To ensure consistency, we assume that the mean risk score for sick patients (*μ<sub>s</sub>*) is greater than that for healthy patients (*μ<sub>h</sub>*), such that *μ<sub>h</sub>* < *μ<sub>s</sub>*. In addition, we assume equal variances (*σ*) for both distributions, following similar assumptions made in prior work (e.g., see refs. 4–6). The equal variance assumption is valid for eight out of the eighteen algorithms derived from the mammography crowdsourcing challenge data used in our empirical analysis. While relaxing this assumption leads to changes in optimal costs, the theoretical insights remain robust. Furthermore, the backtesting experiments in the numerical section do not rely on the equal variance assumption. Under these assumptions, the risk prediction for a specific mammogram is assumed to follow a normal distribution, *N(μ<sub>s</sub>, σ)* for patients with the disease and *N(μ<sub>h</sub>, σ)* for those without it. The performance of an algorithm, in terms of AUC, is then defined as

$$AUC = \Phi\left(\frac{\mu_s - \mu_h}{\sqrt{2}\sigma}\right) = \Phi\left(\frac{I}{\sqrt{2}}\right). \quad (3)$$

In (3), the function *Φ(·)* represents the standard normal cumulative distribution function and *I* :=  $\frac{\mu_s - \mu_h}{\sigma}$  represents the information content of the algorithm<sup>7</sup>. Observe that the information content measure, *I*, has a monotonic relationship with the AUC metric.

The expected cost of a screening decision incurred by a healthcare organization under the delegation strategy, denoted as  $C_D(t_D)$  (see Equation (17)), and the automation strategy, denoted as  $C_A(t_A)$  (see Equation (20)), is determined by the respective thresholds applied in each approach ( $t_D$  for delegation,  $t_A$  for automation). Next, we outline additional factors considered in the cost calculation. Each use of the algorithm incurs a constant cost, denoted as  $c_a$ . This fixed cost per mammogram applies to both per-use pricing models and subscription-based pricing, provided the subscription covers all mammography exams within a predefined capacity limit. The cost of involving a human expert in the decision is denoted as  $c_e$ . In addition, the healthcare organization incurs a cost of  $c_f$  for each follow-up procedure performed when a patient has a suspicious finding. If the physician or AI mistakenly classifies a sick patient as healthy, the organization faces potential litigation risks. We assume that litigation occurs when cancer is missed, meaning a case of cancer is incorrectly classified as negative. This assumption is supported by our review of the breast imaging malpractice literature, which indicates that the majority of litigation stems from false negatives. Specifically, 93% of cases involve delays in cancer diagnosis, where cancer is present but not detected<sup>8,9</sup>. In accordance with empirical findings, we do not consider extra costs related to false positives beyond those incurred from follow-up procedures. We define  $c_l$  as the expected litigation cost for a false negative (FN) outcome. This cost accounts for the probability of a patient initiating legal action due to a missed cancer, the likelihood of the lawsuit being successful, and is therefore referred to as the expected litigation cost. The expected litigation cost also accounts for population heterogeneity in critical factors, such as the occurrence of interval cancers and differences in cancer grades. Full details of our mathematical model are provided in the Methods section.

### Optimal allocation of mammograms between radiologists and AI

Within the context of our framework, the healthcare organization's objective is to reduce the costs associated with mammography-based breast cancer screening using AI. This objective aligns with an increasingly prevalent payment model in the United States: population-based payments for primary care services through capitation arrangements. Under this model, healthcare organizations, such as physician groups specializing in primary care, receive prospective fixed fees for delivering primary care services, including mammography screening and follow-up procedures. Notably, this compensation model separates the volume of services rendered from the reimbursement received, thus promoting cost-effective practices. (see ref. 10 for further details on capitation-based payment models). Due to Civil Monetary Penalties and Anti-Kickback statutes, healthcare organizations are legally prohibited from compensating physicians based on the intensity of services, whether more or less intense. As a result, healthcare organizations cannot financially incentivize physicians to make medical decisions that prioritize profitability. Driven by these considerations, the healthcare organization wants to minimize its total costs without compromising physician autonomy in clinical decision-making. The organization's overall objective,  $C^*$ , is formally defined as:

$$C^* := \min\{C_E, C_D^*, C_A^*\}. \quad (4)$$

That is, the healthcare organization first finds the optimal cost for delegation,  $C_D^* := \min t_D C_D(t_D)$ , and automation,  $C_A^* := \min t_A C_A(t_A)$ , by respectively optimizing  $t_D$  and  $t_A$  for the two strategies involving AI. We make two simplifying assumptions in solving the model, Assumptions 1 and 2, which are technically and practically satisfied, as detailed in the Methods section. The overall optimization involves comparing optimal solutions with the expected cost of the expert-alone strategy,

$C_E$  (see Equation (12)). Theorem 1 in Section S1 presents closed-form solutions for optimal thresholds.

Next, we analytically derive the optimal strategy. Building on insights from the value of information and medical decision-making literature (e.g., see refs. 11,12), the characterization of the optimal strategy within the parameter space results in intuitive thresholds consistent with those commonly observed in these fields. To ensure clarity in the presentation, we introduce the following terms, which are naturally derived from our model. Let

$$I_{ED} := I \in \mathbb{R} \text{ s.t. } C_D^* = C_E, \quad (5)$$

$$I_{EA} := I \in \mathbb{R} \text{ s.t. } C_A^* = C_E, \quad (6)$$

$$I_{DA} := I \in \mathbb{R} \text{ s.t. } C_D^* = C_A^*. \quad (7)$$

$I_{ED}$ ,  $I_{EA}$ , and  $I_{DA}$  denote the values of the algorithm's information content at which the expected costs of the expert-alone and delegation strategies, expert-alone and automation strategies, and delegation and automation strategies, respectively, are the same as each other. Additionally, let:

$$BA := \frac{P_E(TP) + P_E(TN)}{2} \quad (8)$$

$$NFB := P_E(TP)\lambda[c_l - c_f] - P_E(FP)(1 - \lambda)c_f, \quad (9)$$

represent balanced accuracy and the net financial benefit due to expert decisions, respectively. Balanced accuracy (BA) is the average of the sensitivity and specificity of the expert and is equivalent to another related performance measure called Youden Index<sup>13</sup>. The net financial benefit (NFB) is calculated by subtracting the prevalence-adjusted cost of false positives ( $P_E(FP)(1 - \lambda)c_f$ ) from the prevalence-adjusted marginal benefit of preventing litigation risk through true positives ( $P_E(TP)\lambda(c_l - c_f)$ ). Proposition 1 illustrates how the interplay between radiologist accuracy, cost factors, and AI algorithm performance influences the choice of the optimal AI implementation strategy. It defines the conditions under which the expert-alone, delegation, or automation strategies are optimal.

**Proposition 1.** Assume  $I_{ED} < I_{DA}$  so that delegation is a possible alternative. The optimal strategies for various conditions are outlined in Table 1.

The characterization of the optimal strategy is influenced by the comparison between disease prevalence and the relative economic consequences of false-positive versus false-negative decisions, referred to as the cost ratio of false decisions. Specifically, Tables 1a and b present distinct strategies depending on whether disease prevalence is low or high relative to these cost ratios. In both low and high-prevalence cases, the optimal strategy is determined by both the performance of the human expert and the AI algorithm. The columns in Tables 1a and b categorize the expert's performance into two groups, distinguishing between cases where the net financial benefit is either low or high. The rows in Tables 1a divide the expert's net clinical benefit, assessed through balanced accuracy, into categories of low and high values. They also categorize the AI algorithm's discriminative performance in terms of  $I$  into low, medium, or high values (Tables 1a and b). This framework extends the foundational work of<sup>41</sup> on decision curve analysis, which evaluates the utility and economic value of prediction-based decisions by weighing their benefits against their harms. It applies these principles to combining human and machine predictions within a workflow and subsequent decisions.

Proposition 1 offers valuable insights. First, disease prevalence plays a pivotal role in determining the optimal strategy. Let  $\lambda' = \frac{c_f}{c_l}$

**Table 1 | Optimal strategy**

(a) $\lambda < c_f/c_l$			
		$NFB < c_e - c_a$	$(c_e - c_a) \leq NFB$
$BA < 0.5 + \frac{c_e c_l}{2c_f(c_l - c_f)}$	$I \leq I_{EA}$	A	E
	$I_{EA} < I$		A
$0.5 + \frac{c_e c_l}{2c_f(c_l - c_f)} \leq BA$	$I \leq I_{ED}$	D	E
	$I_{ED} < I \leq I_{DA}$		D
	$I_{DA} < I$	A	A
(b) $\lambda \geq c_f/c_l$			
		$NFB < (c_e - c_a) + \lambda c_l - c_f$	$(c_e - c_a) + \lambda c_l - c_f \leq NFB$
$I \leq I_{ED}$	A		E
$I_{ED} < I \leq I_{DA}$	A		D
$I_{DA} < I$	A		A

represent the critical prevalence value, defined as the ratio of the cost of a false positive to that of a false negative (the cost ratio). Disease prevalence is considered high when  $\lambda \geq \lambda'$  and low otherwise. Our findings indicate that the level of disease prevalence significantly influences the healthcare organization's optimal strategy. Since breast cancer prevalence varies across populations (e.g., due to differing demographics), these results suggest that healthcare organizations should tailor their strategies to the specific prevalence levels within the patient populations they serve. A broader implication is that the choice among automation, delegation, or expert-alone strategies for different diseases will also depend on disease prevalence. The critical prevalence threshold for determining the optimal strategy will be dictated by the ratio of economic consequences associated with false-positive detections (e.g., unnecessary follow-up costs) versus false-negative errors (e.g., litigation costs from missed diagnoses).

Second, when the net financial benefit due to the decision of the human expert (i.e., the radiologist) is low, automation emerges as the preferred strategy in almost all cases. The sole exception occurs when certain mammography tasks are delegated to human experts, driven by a high clinical benefit that compensates for the low financial return. This scenario arises when the information content of the algorithm is low, and disease prevalence is also low.

Third, when both the net financial and clinical benefits of the expert's performance are sufficiently high, any of the three strategies—expert-alone, delegation, or automation—can be optimal, for both low and high disease prevalence cases. In this scenario, the performance of the algorithm primarily determines the optimal strategy. Both tables illustrate a clear progression. When the algorithm's performance changes from low to medium, the optimal strategy transitions from expert alone to delegation. Similarly, when the algorithm's performance changes from medium to high, the optimal strategy becomes full automation. This progression indicates a transition from the expert-alone approach to the delegation strategy, where both humans and machines share the task of mammography interpretation, and eventually to full automation when the algorithm's performance is sufficiently high.

Fourth, when the disease prevalence and net clinical benefit is low, the increasing performance of algorithms precipitates an abrupt transition from reliance on expert judgment to automated systems. This change arises from greater ambiguity in distinguishing between sick and healthy patients. In such situations, either expert alone or automation is preferred over the delegation strategy.

Finally, when the net financial benefit of the human expert is high, and the performance of the AI is low, the expert-alone strategy becomes the optimal strategy in both low- and high-prevalence scenarios.

As AI technology continues to progress, achieving higher levels of accuracy, efficiency, and reliability, its impact on decision-making processes warrants deeper exploration. In this regard, we present a corollary to Proposition 1, which provides additional insights into the conditions under which automation should be prioritized over human involvement.

**Corollary 1.** For any given  $P_E(TP)$  and  $P_E(FP)$ , there exists a performance level  $I_A$  such that when  $I_A < I$ , the optimal strategy is automation.

Corollary 1 implies that once AI algorithms surpass a specific performance level, automation becomes the optimal strategy for mammography-based breast cancer screening. This shift to automation signifies a critical performance point beyond which the reliance on human radiologists are minimal. After a tipping point of AI performance, healthcare organizations can realize significant efficiency gains through the use of AI algorithms to replace some human tasks. We also note that the critical performance threshold required to switch to automation,  $I_A$ , is higher for those tasks where expert performance is higher. Hence, in areas where human experts perform particularly well, the path to full automation may be slower, requiring AI to reach even greater levels of performance to become the preferred option.

### Asymmetric litigation costs for human vs. machine liability

An important cost dynamic we capture in our model is the economic consequences of missing cancer cases, the case of a false-negative, and the ensuing litigation scenario. In this subsection, we analyze the situation where litigation costs differ between the case when an AI algorithm makes an incorrect decision and one in which a human makes an incorrect decision. The current regulatory environment in the U.S. does not provide a definitive answer to this question in either the health or other contexts<sup>14</sup>. In establishing liability for a missed cancer case by a doctor, we ask whether a reasonably competent and skilled healthcare professional would have made the same mistake. That is the basic definition of the medical standard of care. As machines are commercial products, they may be subject to product liability law, a potentially stricter standard, which may result in greater liability costs. Hence, in this section, we analyze the scenario in which this premise holds true: the liability costs for machines are higher than the same for humans.

We modify our model to study the case in which human vs. machine liability are different. In the next proposition, we study the effect of having a larger litigation cost for the algorithm when a sick patient is deemed as healthy as compared with the litigation cost for the human expert on optimal strategy.

**Proposition 2.** A higher liability for machines could alter providers' preferences away from automation (or delegation) to expert-alone solutions.

Proposition 2 suggests that holding machines to a stricter standard as compared to humans could encourage increased human involvement in prediction based decisions. This result, although intuitive, has important implications for the future of work and points to a significant policy lever that social planners can utilize. First, even the best-performing algorithms can produce imperfect predictions; hence, liability cases will be unavoidable when we let machines make decisions. Algorithms can even perform strictly better than humans do on average. However, they may lead to unintended consequences or adverse effects. For example, in healthcare, algorithms could miss clinical context and data outside clinical records, which humans may not necessarily miss. In such a scenario, human involvement may be preferable to a better-performing algorithm. One way to induce such preference of humans or society could be holding machines to a stricter standard (by imposing larger litigation costs for algorithms). Second, social planners who regulate liability cases can set differential standards for humans or machines for a healthier transition to human-



**Table 2 | Parameter Estimates**

Parameters	Comments	Estimates
$\lambda$	The Cancer Statistics Center website <sup>17</sup>	$126.9 \times 10^{-5}$
$C_l$	Inflation-adjusted low/high compensation paid <sup>18</sup>	\$756,156/\$1,157,866
$C_f$	Based on use patterns in the Group Health BCSC registry <sup>19</sup>	[\$1223.90, 2612.46\$]
$C_e$	Based on Medicare reimbursement rates <sup>19</sup>	\$167.04
$C_a$	Total cash prize per mammogram in DM Challenge <sup>2</sup>	\$4.47
$P_E(TP)$	Radiologist sensitivity in the DREAM challenge data <sup>2</sup>	0.856
$P_E(FP)$	Radiologist false positive rate in the DREAM challenge data <sup>2</sup>	0.095

Notes. These estimates are derived from the indicated references, which provide only the mean values.

machine combined solutions and automation across firms. Social planners can use liability costs as a policy lever to account for factors that a cost-minimizing firm would otherwise may not. For example, the broader good of society could require a slower immersion of AI, especially in life-and-death decisions. Developments in AI technologies will likely outpace the legal developments on how we should hold algorithms accountable is complex, dynamic, and ever evolving. Until we learn more about how to regulate AI and prevent injuries, a stricter standard and costs on algorithmic liability can help with the pacing problem (e.g., see ref. 15 for a discussion on regulating AI in the context of negligence laws). The differential treatment of human vs. machine liability can help with the pacing problem between technological and legal developments.

**Empirically studying AI-based allocation of mammograms**

We parameterize our model using multiple data sources and conduct two types of numerical experiments. In the first set of numerical experiments, we use estimates from large-scale, multi-site US mammography studies, as summarized in Table 2 along with their references. These parameters and our analytical model results help characterize optimal strategies across varying human and AI performance. In the second set of numerical experiments, we use algorithmic predictions and radiologist assessments from the Digital Mammography DREAM (Dialog on Reverse Engineering Assessment and Methods) Challenge<sup>2</sup>. The challenge was organized by leading AI and healthcare organizations including IBM Research, Icahn School of Medicine at Mount Sinai, Sage Bionetworks, and Kaiser Permanente Washington. The challenge proposed a crowdsourcing contest that aimed to develop AI algorithms for mammography-based breast cancer detection. The algorithms assessed risk using a continuous scale from zero to one, representing the probability of cancer. More details about the challenge can be found online at the challenge site<sup>16</sup>. Using the DREAM data and parameter estimates from the medical literature, we compute optimal thresholds for each strategy and identify the one with the lowest expected cost (through numerical search), independent of the assumptions made in the analytical model.

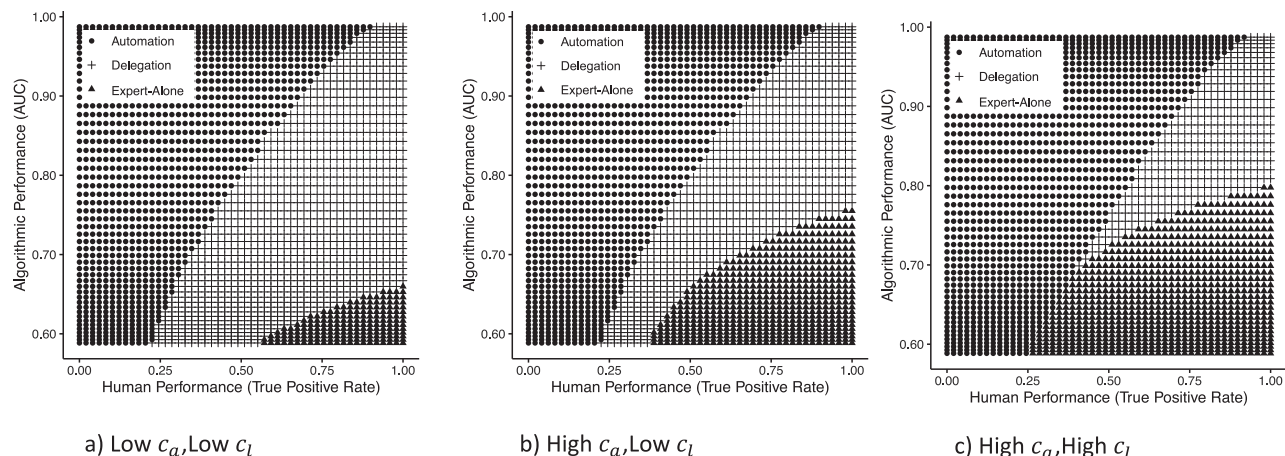
We have proprietary access to the predicted scores of AI algorithms, the true outcome of whether or not the patient developed cancer within a year from screening, and the radiologist’s assessment of the mammogram (benign or recall for further study). We report the radiologists’ performance in Table 2. In conducting our experiments, we used the top-ranked algorithms, which were offered \$1, 000, 000 if they can beat radiologists; they were eventually rewarded a total of \$140,000 cash prize (no algorithm was able to meet the initial goal of beating radiologists). We obtain the per-mammogram algorithm cost leveraging the total cash prize offered and eventually given (based on an assumption that the designers chose the award in a way to implicitly price these algorithms). Specifically, we assume  $c_a$  to range between \$140, 000/25, 657 = \$5.46 (\$5.96 in 2020 dollars) and \$1, 000, 000/25, 657 = \$38.98 (\$42.59 in 2020 dollars), which we consider as the valuation of the AI systems by domain experts.

**Numerical characterization of optimal strategy.** We start by characterizing the optimal strategy as a function of cost and performance parameters. For populating model parameters, we use point estimates as presented in Table 2 and numerically demonstrate our theoretical findings.

In Fig. 1, we depict (i) how the apportioning of work between expert radiologists and algorithms depends on the performance differential between the two and (ii) the impact of algorithm and liability costs on how the work is shared between them. We fix the false-positive rate for radiologist performance to that in the Dream Challenge data and vary their true-positive rate. The figure suggests that for a fixed human performance level that is not too large, automation is the preferred strategy when the algorithm performance exceeds a certain AUC performance threshold. This threshold is increasing in the human performance. In contrast, expert alone is the preferred strategy when human performance is very high while the algorithm performance is low enough. When neither the algorithm nor the radiologist dominates the other in terms of their performance, delegation could be the preferred strategy. This high-level characterization of optimal strategy lays out a foundation for how disparate cost and performance parameters affect the healthcare organization’s strategy choices.

A comparison of sub-figures reveals additional insights. Figure 1a and b suggest that a reduction in algorithmic costs replaces some of the parameter regions of the expert-alone strategy with the delegation strategy. In other words, reducing algorithmic costs facilitates human-machine combined solutions in lieu of human-alone approaches. A change in litigation costs, based on the comparison of Fig. 1b and c, has a similar yet broader effect. Lower litigation costs, when compared to higher litigation costs, have the effect of increasing the region in which delegation is optimal while reducing the regions of optimality for both automation as well as expert-alone strategy. These interesting results have a two-fold explanation. With low litigation costs, in the region with higher human performance, an expert’s (higher) cost becomes more ‘affordable’ and thereby causes some regions of automation to switch to delegation. On the other hand, the expert-alone area shrinks due to a different reason. The lower litigation costs increases the tolerance for false negatives. This means the optimal threshold under a delegation strategy increases resulting in lower reliance on the expert under a delegation strategy. It also means that delegation becomes a more cost-effective strategy against experts alone in some regions. The result around the impact of lower litigation costs on human involvement implies that policies such as placing a cap on damages (that a judge can award in a litigation case) through tort reforms can impact how the work is shared between humans and machines. Also, observe that when both the algorithm’s and human’s performance are low enough while the algorithm and litigation costs are high, the delegation strategy may not be feasible (see the lower left region in Fig. 1c). The observation suggests that, when costs are high, the feasibility of delegation strategy requires high human and algorithm performances.

**Back testing our model.** In this subsection, we validate our proposed approach by retrospectively testing the cost/performance of



**Fig. 1 | The figure shows three optimal strategy regions—expert-alone, delegation, and automation—based on the relative performance of AI and the radiologist. The x-axis represents the AUC of AI algorithm performance, while the y-axis represents the radiologist's true positive rate. Panel (a) highlights how lower**

AI and litigation costs impact the preference of three strategies. Using Panel (a) as a benchmark, Panel (b) illustrates how higher algorithm costs shift strategy preferences, while Panel (c) examines the combined impact of increased algorithm and litigation costs on strategy selection. Source data are provided as a Source Data file.

different AI strategies based on real data. We have algorithm risk scores for 25,657 distinct mammograms in the holdout sample produced by 18 challenge contestants. The dataset uses 2D digital screening mammograms interpreted by a single radiologist. This sample had a breast cancer prevalence of 1.10%, a value substantially larger than the 0.127% incidence reported for the broader population<sup>17</sup>. The challenge designers chose to create a sample with a higher than typical cancer rate for better assessing algorithm performance, especially for mammograms indicating cancer. In conducting our experiments, we undersample in a way to match the population prevalence and repeat this sampling procedure 100 times; the undersampling eliminates the systematic bias in cost estimates due to the high incidence in the holdout sample. For each iteration of the sampling procedure, we evaluate the three strategies that our problem considers. In assessing the expected costs for the expert-alone strategy, we obtain varying costs due to changing radiologist performance across samples. We calculate the mean and standard deviation of costs to get a distribution. In assessing the delegation and automation strategies, we conduct an exhaustive search over  $t_D$  and  $t_A$ , respectively. Dividing the risk score range into 1000 equal intervals for each of the algorithms, we find the computationally optimal thresholds for all samples. We obtain the means and standard deviations of costs for the selected thresholds across samples and use it to calculate the expected costs for the delegation and automation strategies. For each of the 18 algorithms, we compare the average costs associated with each strategy using a two-sided t-test and assess optimality at a 95% confidence level. The null hypothesis is that the average costs of compared strategies are equal. If the estimated costs are statistically indifferent across a pair, we prefer expert alone, delegation, and automation as the optimal strategy in that order. Note that the backtesting experiments do not rely on the assumptions presented in the modeling section.

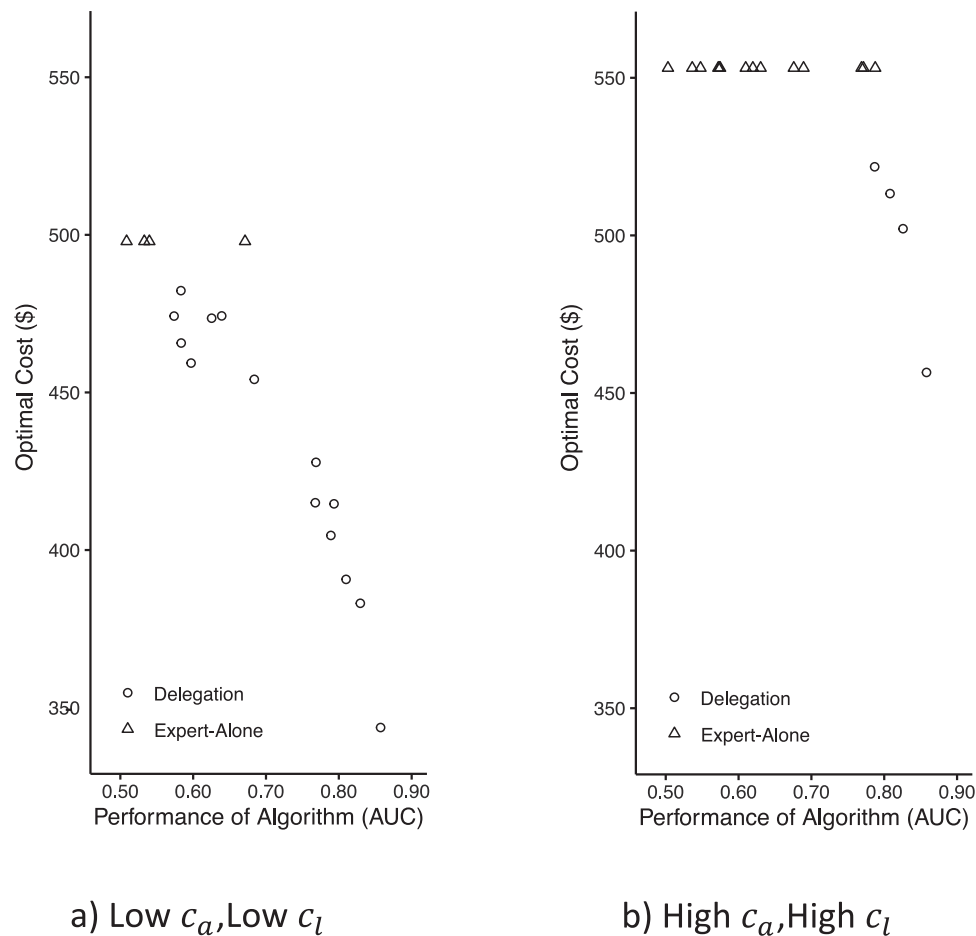
Figure 2 quantifies the mean cost estimates corresponding to the (computationally) optimal strategies for each of the submitted algorithms (we assume algorithm cost to be the same for all algorithms). When both the algorithm and litigation costs are low as in Fig. 2a, the optimal strategy is delegation for most algorithms even when their performance is lower. The difference in optimal costs between using higher and lower-performing algorithms in the redesign is large (notice the steep reduction in costs as the algorithm performance increases). The expected costs under the delegation strategy using the top-performing algorithm is \$343.69 which suggests a 30.1% reduction

from the expert-alone strategy with an associated cost of \$497.92. When both the algorithm and litigation costs are high, as in Fig. 2b, we observe only the top-performing algorithms to become a part of the mammography operations (i.e., delegation is optimal only for the top four algorithms with AUC values ranging from 0.789–0.857). The delegation strategy is optimal for the best-performing algorithm (i.e., AUC = 0.857 for  $A_1$ ). The expected costs under the delegation strategy is \$456.53 which suggests a 17.5% reduction from the expert-alone strategy with an associated cost of \$553.10. The reductions in both cases, low/high algorithm, and liability costs suggest that potential savings from the delegation strategy could be substantial to healthcare organizations. Also, it could be especially valuable for society considering the 40 million mammograms conducted every year, an estimate from the U.S. Food and Drug Administration.

## Discussion

This research examines how a healthcare organization strategically allocates tasks between human experts and AI systems. The organization evaluates the adoption of an AI algorithm to automate some or all post-mammography decisions, aiming to lower operational costs. It considers three strategies: relying exclusively on human expertise (expert alone), combining human input with machine assistance (delegation), or fully automating decisions with AI (automation). We use a probabilistic model, grounded in statistical principles of machine learning, to represent algorithmic decisions. By applying this model, we minimize costs and identify conditions under which each strategy becomes optimal. We also characterize these strategies based on the cost and performance of human experts and AI algorithm. To quantify the economic outcomes, we incorporate estimates from medical studies and results from a crowdsourced competition on AI-based breast cancer detection.

Our study offers valuable guidance to healthcare organizations on integrating AI into their workflows by balancing tasks between expert radiologists and algorithms. A notable insight from our findings highlights the role of disease prevalence as a key factor in determining AI's optimal use. Specifically, the balance between high and low prevalence, combined with the cost trade-offs of false positives and false negatives, can shift the preference toward either a delegation strategy (human-machine combination) or full automation. This observation suggests that applying a uniform approach across all diseases may not be effective. Instead, organizations should tailor their AI adoption and operational redesign decisions to the specific characteristics of each disease.



**Fig. 2 |** The figure presents mean cost estimates from backtesting experiments, comparing preferred strategies and associated costs across 18 AI algorithms submitted in a crowdsourced data competition trained on real-life mammograms. The x-axis represents algorithm AUC values, and the y-axis shows the expected cost per mammogram in USD. Panel (a) depicts results with low algorithm

and litigation costs, where delegating to the top-performing AI reduces costs by 30.1% compared to the expert-alone strategy. Panel (b) shows that when both costs are high, delegation is limited to the top four algorithms, resulting in a smaller 17.5% cost reduction. Source data are provided as a Source Data file.

The costs associated with algorithms and false assessments, such as litigation expenses from false negatives, play a crucial role in shaping a healthcare organization's optimal strategy. When AI demonstrates high predictive performance, the organization may assign full responsibility to either the AI or the human expert, depending on the cost of implementing the algorithm. In such cases, reliance on human-only or machine-only strategies often eliminates the delegation approach. Similarly, higher litigation costs can drive this effect. Automation may become more appealing than delegation, depending on how AI performance compares to that of human experts. However, differences in litigation costs between algorithms and humans may induce the uptake of hybrid strategies where human involvement remains part of the decision-making process.

The numerical experiments highlight both the practicality and benefits of the proposed approach. Based on the current capabilities of AI in mammography and existing clinical practices, our findings indicate that a human-machine combined workflow (the delegation strategy) emerges as the most effective option. This strategy could achieve significant cost reductions, ranging from 17.5% to 31.1%. By leveraging AI to assign specific tasks to human experts while automating others, healthcare organizations can unlock substantial efficiencies and cost savings.

We would like to note that our model reflects U.S. screening practices. In non-U.S. contexts, such as Europe where double reading is

standard, alternative strategies should be considered. For instance, AI could serve as a second reader alongside a radiologist reducing the need for two radiologists to review and reach a consensus, potentially offering cost savings. Given that the performance characteristics of AI combined with a radiologist and the cost considerations, including litigation, differ in European contexts, exploring alternative workflow strategies and parameterizations would be a valuable direction for future research.

In conclusion, this paper introduces an operational framework for integrating AI into routine clinical decision-making, with a focus on mammography-based breast imaging. We develop a mathematical characterization of the problem and validate our theoretical insights using data from a real-world crowdsourcing competition on breast cancer prediction from mammograms. While the study uses mammography as a case example, the framework is adaptable to various clinical scenarios—such as pathology and dermatology—where predictive accuracy is essential, and AI holds significant potential to improve efficiency.

## Methods

### Details of mathematical model

We define the expected costs associated with each strategy as follows:

$$C_E := c_e + \overbrace{P(\eta=s)c_f}^{\text{Expected follow-up costs}} + \overbrace{P(\eta=h|s)\lambda c_l}^{\text{Expected litigation costs}} \quad (10)$$

$$= c_e + \lambda c_l + P_E(TP)\lambda[c_f - c_l] \quad (11)$$

$$+ P_E(FP)(1 - \lambda)[c_f], \quad (12)$$

$$C_D(t_D) := c_a + \overbrace{P(r \geq t_D)c_e}^{\text{Expected expert costs}} + \overbrace{P(r \geq t_D, \eta = s)c_f}^{\text{Expected follow-up costs}} \quad (13)$$

$$+ \overbrace{P(r < t_D | s)\lambda c_l + P(r \geq t_D, \eta = h | s)\lambda c_l}^{\text{Expected litigation costs}} \quad (14)$$

$$= c_a + \lambda c_l \quad (15)$$

$$+ P(r \geq t_D | s)\lambda[c_e + P(\eta = s | s)(c_f - c_l)] \quad (16)$$

$$+ P(r \geq t_D | h)(1 - \lambda)[c_e + P(\eta = s | h)c_f], \quad (17)$$

$$C_A(t_A) := c_a + \overbrace{P(r \geq t_A)c_f}^{\text{Expected follow-up costs}} + \overbrace{P(r < t_A | s)\lambda c_l}^{\text{Expected litigation costs}} \quad (18)$$

$$= c_a + \lambda c_l + P(r \geq t_A | s)\lambda(c_f - c_l) \quad (19)$$

$$+ P(r \geq t_A | h)(1 - \lambda)c_f. \quad (20)$$

The healthcare organization's overall objective  $C'$  is the following:

$$C^* := \min \left\{ C_E, \min_{t_D} C_D(t_D), \min_{t_A} C_A(t_A) \right\}. \quad (21)$$

In our optimization model, we rely on two technical assumptions that underpin our analysis. Although technical in nature, these assumptions are practically met in the context of breast screening, as explained below.

**Assumption 1.**  $c_a/(1 - \lambda) < c_e < c_f < c_l$ .

Assumption 1 describes the relationships between various costs. According to this, liability costs for a false negative,  $c_l$ , are the highest. The cost of a follow-up exceeds that of an expert's evaluation of a mammogram. These assumptions align with observed costs in practice. For instance, the average settlement for a litigation claim was reported as \$485,348 (\$756,156 adjusted for 3% inflation to 2020)<sup>18</sup>. Similarly, the cost of a digital screening mammogram read by an expert radiologist averaged 139.89 (167.04 adjusted for 2020), while diagnostic follow-up costs (e.g., additional imaging or biopsies) ranged between \$1,025.00 and \$2,187.89 (\$1,223.90 to \$2,612.46 adjusted for 2020)<sup>19</sup>. Finally, the leftmost inequality indicates that the cost of using the AI algorithm, divided by  $1 - \lambda$ , is lower than the cost of a human expert. Since  $1 - \lambda < 1$ , it follows that  $c_a < \frac{c_e}{1 - \lambda}$ , and consequently,  $c_a < c_e$ , meaning the algorithm is less expensive than a human expert. It is important to note that although the leftmost inequality is primarily technical (as it involves division by  $1 - \lambda$ ), for all cancers, including breast cancer, the population prevalence  $\lambda$  is extremely low (e.g.,  $\lambda = 126.9 \times 10^{-5}$  for breast cancer<sup>17</sup>), resulting in  $1 - \lambda$  being nearly equal to 1.

**Assumption 2.**  $P_E(TP) > \frac{c_e}{c_l - c_f}$  and  $P_E(FP) < \frac{c_f - c_e}{c_f}$ .

Assumption 2 provides a lower bound for the true-positive probability and an upper bound for the false-positive probability of the human expert. Accordingly, we assume that the human expert exhibits

high true-positive rates and sufficiently low false-positive rates, ensuring accurate identification of diseased cases while minimizing unnecessary follow-up procedures. These bounds are technical and serve to exclude the unrealistic scenario where the expert's performance adds no more value than random guessing of disease status. According to a recent study, radiologists in clinical practice achieve a sensitivity (true-positive rate) of 73% and a specificity (false-positive rate of 4%)<sup>20</sup>. Based on the previously mentioned cost figures, the conditions  $P_E(TP) \gg 0.01\%$  and  $P_E(FP) \ll 86.3\%$  are satisfied, providing practical support for this assumption.

Consistent with prior machine learning research in mammography, we assume conditional independence between the algorithm's risk score and the human expert's predictions given the patient's health status. This assumption implies that, conditional on cancer status, the AI and radiologist make independent predictions and may not necessarily identify the same cases. As a result, this assumption highlights the complementary diagnostic strengths of AI and radiologists. It is important to note that the backtesting experiments in the numerical section utilize AI predictions from the DREAM challenge without imposing the analytical model's assumptions. In addition, we conduct an exhaustive search of over-optimization parameters across different strategies, ensuring that the numerical results remain independent of these assumptions.

## Ethics

Sage Bionetworks obtained IRB approval to conduct the Digital Mammography Challenge and to share the best predictive models and accompanying methods for research use. The IRB also granted a full waiver of authorization under HIPAA to enable access and analysis of the mammogram images and survey data to Sage Bionetworks and partner challenge organizers. The justification for granting the waiver of authorization under HIPAA is that: 1- The images and survey data were already collected and were available to use for research purposes under an opt-out mechanism (risk factor questionnaire) or under a waiver of informed consent (de-identified mammogram images and BIRAD scores). 2- Sage Bionetworks, the sponsor, and other Challenge organizers didn't know the subjects' identities. It is impossible to re-contact the data subjects to obtain authorization without breaching their confidentiality, which poses a greater risk to the data subjects than having their images analyzed. 3- The project is expected to analyze about 100,000 digital images. It was impractical, time-consuming, and prohibitively costly to re-contact the large number of people whose images would be analyzed in this Challenge.

## Datasets

This study leverages two primary datasets to support its analysis. The first dataset is synthetic data designed to simulate a controlled environment for validating theoretical models. This dataset includes varying values for human and algorithmic performance, enabling a comprehensive analysis of how the optimal strategy evolves with changes in these performance levels. The synthetic dataset can be reproduced using the code available in our GitHub repository<sup>21</sup>.

The second dataset originates from the Digital Mammography DREAM Challenge, a real-world crowdsourcing competition. It includes breast cancer risk predictions from eighteen algorithms, radiologist assessments, and cancer outcomes (cancer present or not) for 25,657 patients used in the validation phase of the challenge (further details about the challenge are available in ref. 2). The algorithms generated risk scores based solely on mammography images, without incorporating any additional clinical or demographic information. We use the risk predictions by algorithms, radiologist assessments, and cancer outcomes in their original form, without additional processing.



## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The synthetic data used to create Fig. 1 and the associated results can be generated using the code in our GitHub repository, hosted at [https://github.com/ahsen1402/delegation\\_project](https://github.com/ahsen1402/delegation_project)<sup>21</sup>. The Digital Mammography DREAM Challenge data used to create Fig. 2 and associated results are available under restricted access due to its sensitive nature. The data use agreement that allowed us to use the data for the DREAM Challenge explicitly states that the data cannot be redistributed in total or in parts. The data can be accessed through a request to the data provider, Kaiser Permanente (KP) Washington. Access to KP Research Data resources by non-KP researchers requires collaboration with a KP researcher, who can help refine the phenotype of interest. The KP Research Bank hosts data and analytic tools in a secure analytic platform that is managed by KP. Researchers will be given access to this platform to perform analyses. The website to apply for access is <https://researchbank.kaiserpermanente.org/for-researchers/apply-for-access/>. Source data are provided with this paper.

## Code availability

The analyses reported in this study used the statistical software R version 4.4.2. We also used R package ggplot2 version 3.5.1 and prisma version 2.4.4 for creating visualizations. The entire codebase utilized in this manuscript is publicly available under the Apache License on the following GitHub repository: [https://github.com/ahsen1402/delegation\\_project](https://github.com/ahsen1402/delegation_project)<sup>21</sup>.

## References

- Lee, C. I. & Elmore, J. G. Artificial intelligence for breast cancer imaging: the new frontier? *J. Natl. Cancer Inst.* **111**, 875–876 (2019).
- Schaffter, T. et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Network Open* **3**, 200265–200265 (2020).
- Lång, K. et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (masai): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24**, 936–944 (2023).
- Fong, H., Kumar, V., Mehrotra, A., Vishnoi, N.K.: Fairness for auc via feature augmentation. Preprint at <https://doi.org/10.48550/arXiv.2111.12823> (2021).
- Ahsen, M. E., Ayvaci, M. U. S. & Raghunathan, S. When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Inf. Syst. Res.* **30**, 97–116 (2019).
- Ayvaci, M. U. S., Ahsen, M. E., Raghunathan, S. & Gharibi, Z. Timing the use of breast cancer risk information in biopsy decision-making. *Prod. Oper. Manag.* **26**, 1333–1358 (2017).
- Metz, C.E. Basic principles of roc analysis. *Semin. Nucl. Med.* **8**, 283–298 (1978).
- Murphy, B. L. et al. Breast cancer litigation in the 21st century. *Ann. Surg. Oncol.* **25**, 2939–2947 (2018).
- Brady, A. P. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging* **8**, 171–182 (2017).
- James, B. C. & Poulsen, G. P. The case for capitation. *Harv. Bus. Rev.* **94**, 102–11 (2016).
- Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).
- Felder, S., Mayrhofer, T. *Medical Decision Making*. Third edition ed. Springer, Heidelberg Platz 3, 14197 Berlin, Germany (2017).
- Fluss, R., Faraggi, D. & Reiser, B. Estimation of the Youden Index and its associated cutoff point. *Biom. J.* **47**, 458–472 (2005).
- Abbott, R. *The Reasonable Robot: Artificial Intelligence and the Law*, pp. 1–17. (Cambridge University Press, Cambridge 2020).
- Selbst, A. D. Negligence and AI's human users. *BUL. Rev.* **100**, 1315 (2020).
- Sage Bionetworks: Digital Mammography DREAM Challenge. <https://www.synapse.org/#!Synapse:syn4224222/wiki/> (2018).
- American Cancer Society: Cancer Statistics, Springfield, IL. (2022).
- Studdert, D. M. et al. Claims, errors, and compensation payments in medical malpractice litigation. *N. Engl. J. Med.* **354**, 2024–2033 (2006).
- Stout, N.K. et al. Benefits, harms, and costs for breast cancer screening after us implementation of digital mammography. *J. Natl. Cancer Inst.* **106**, <https://doi.org/10.1093/jnci/dju092> (2014).
- Salim, M., Dembrower, K., Eklund, M., Lindholm, P. & Strand, F. Range of radiologist performance in a population-based screening cohort of 1 million digital mammography examinations. *Radiology* **297**, 33–39 (2020).
- Ahsen, M.E. Economics of AI and Human Task Sharing for Decision-Making in Screening Mammography. <https://doi.org/10.5281/zenodo.14269705> (2024).

## Acknowledgements

We acknowledge Natasha Stoloitzky-Brunner for her contribution to creating the featured image for our paper.

## Author contributions

M.E.A.: Conception and design of study, analysis, and interpretation of data; drafting portions of the first draft and taking part in subsequent revisions; final approval of the submitted manuscript; agreement to be accountable for the accuracy and integrity of the work M.U.S.A.: Conception and Design of study; analysis and interpretation of data; drafting portions of the first draft and taking part in subsequent revisions; final approval of the submitted manuscript; agreement to accountable for the accuracy and integrity of the work R.M.: Interpretation of data; critical revision of drafts for important intellectual content; final approval of the submitted manuscript; agreement to accountable for the accuracy and integrity of the work G.S.: Interpretation of data; critical revision of drafts for important intellectual content; final approval of submitted manuscript; agreement to accountable for the accuracy and integrity of the work.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-57409-1>.

**Correspondence** and requests for materials should be addressed to Mehmet Eren Ahsen.

**Peer review information** *Nature Communications* thanks Nathaniel Hendrix, Nisha Sharma, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025