

Economics of AI and Human Task Sharing for Decision-Making in Screening Mammography

Corresponding Author: Professor Mehmet Eren Ahsen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for the opportunity to review this manuscript, which seeks to provide generalizable rules for selecting breast cancer screening strategies using AI and/or radiologist expertise. The paper is generally clearly written. I especially appreciated the exploration of asymmetrical liability, for which the authors provided a nice motivation. The use of DREAM Challenge data was also helpful, and I appreciated the author's point that their work might allow for better personalization of strategies based on risk.

There are a few points where readers might benefit from more explanation or a more thorough exploration of assumptions.

1) Material and Methods / Analytic Model: You write: "...the delegation strategy in which humans and machines collaborate in decision-making." This doesn't seem like an accurate description of the delegation strategy, since the radiologist's determination is (at least theoretically) independent of the AI's.

2) Please provide more information about your assumptions on litigation cost. In the supplementary material, you seem to multiply litigation cost by the prevalence of disease, which I don't understand.

3) The assumption of equal variances in the mean score of sick and healthy patients seems worthy of more exploration and/or validation. Does the DREAM Challenge data allow you to validate that assumption? If it does and the assumption proves unreasonable, I'd encourage you to substitute a more data-driven parameter for the variances or to find a way of estimating how sensitive your results are to differences in variance.

4) In Table 1a, the optimal strategy is defined by balanced accuracy (BA) and net financial benefit. Could you provide an intuition for how you arrived at the BA cutoff used in the table?

5) Typo in Table 2 ("Medicate" instead of "Medicare").

6) Does your work rely on an assumption that AI and radiologists would detect the same set of cases? It seems like a delegation strategy might be cost-minimizing for a low sensitivity AI if it were to detect only the cases that are most likely to be missed by radiologists, for example. Making that assumption explicit (or correcting my intuition) would be helpful.

Reviewer #2

(Remarks to the Author)

The paper presents the economics of incorporating AI fully or partially into breast screening, primarily considering a US healthcare environment where single reading and litigation are more prominent compared to other geographic jurisdictions. This is an important topic when it comes to integration and deploying AI. The paper's significance lies in the design of the economic model presented and providing insight into the parameters that are important for healthcare institutions to consider when applying the model to their situation to understand how to optimally integrate AI into their system in a way that will work for them and meet their needs.

The model designed is presented very well and organised in a way that is easy to follow. It would be worth noting how the USA healthcare environment has steered the design of the model and parameters used, and what adjustments should be considered for other geographies. For example, the strategy where AI replaces the human expert could occur in geographies outside of the USA, however, the strategy where AI replaces only 1 of 2 human experts in geographies where double reading is standard is not considered.

It's important to also note that use of AI performance data from the DREAM challenge may now be outdated with the top algorithm demonstrating a 0.857 AUC. Current day AI algorithms in screening mammography can reach AUCs around 0.95. Some medical device intended use indications even require a minimum of a 0.95 AUC for regulatory clearance in the USA. A higher AUC will have an impact on reducing false positives and false negatives and the level of automatic possible with AI. It would be important and interesting to consider the impact of this higher AI performance bound on the economic model.

Reviewer #3

(Remarks to the Author)

I enjoyed reading this well written paper but I have some questions

1. What model of film reading did the authors use - single or double reading
2. Did the authors consider the expertise of the readers as the recall regarding a breast radiologist will be lower than a non breast radiologist
3. do the authors know the rates of interval cancers - this gives an idea of the cancers missed by human readers and how does this translate for the algorithm as this will impact on litigation costs
4. The authors talked about missing cancers but another caveat is the grade of cancer missed - missing grade 1 versus grade 3 - I am not sure if that can be factored in
5. how did the algorithms assess risk ? was it binary or scale of 0-100 and how was the threshold set

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

As requested by the editorial team, I focused on the proofs presented in the Appendix. The presented model is straightforward and not very involved. From the mathematical point of view, the proofs are on a level which a talented high school student can accomplish, or, alternatively, everyone who has passed an introductory calculus course. For the topics under investigation this methodology is appropriate.

However, the style of presentation is very poor. I have severe doubts that any of the authors has proofread the appendix carefully. Besides subtle issues in the reasoning (which can be easily corrected) the presentation lacks structure, the notation is inappropriate at important spots and basic terminology is used incorrectly. Arguments are partially wrong or insufficient.

In its current form the appendix is clearly unacceptable for publication and requires major revisions before it can be reevaluated.

Below I list some of the issues I encountered when reading the supplement. Since the authors have neither included line numbers nor have they numbered all equations, it is challenging to refer to a precise spot in the text.

(1) Many of the proofs refer to "lemma 1" and "lemma 2". However, there exist no statements which are labeled as "lemma 1" and "lemma 2". Of course, I can guess from the context, which statements the authors have in mind, when referring to "lemma 1" and "lemma 2", but for the sake of readability please label the respective statements proven on pp. 7 as "Lemma 1" and "Lemma 2".

(2) Please use one of the established standard notations to denote derivatives. It is unacceptable to denote the derivative of $C(t)$ with respect to t as $C(t)t$ (see eq. 7 for example) because this looks as the multiplication of $C(t)$ and t . Use either $C'(t)$ or $dC(t)/dt$. The same applies to the last equation on the bottom of p. 7 and to many other equations (e.g., eq. 10, eq 22, unnumbered equation at the bottom of p. 19, eq 27)

(3) The statement of a Proposition or Lemma should mention all assumptions which are required and not only the conclusion. It is not acceptable to present assumptions between the lines and use them e.g. ten or more pages later without clearly stating it. Please adhere to the habits from the field, as they are taught during undergraduate calculus courses.

(4) The third equation on p.8, i.e., the equation without number which follows equation (7) is clearly incorrect. For $t=0$ the left hand-side equals 0 whereas the right hand-side is different from zero. I guess instead of " $(....) t$ " the authors mean " $d (...)/dt$ " (i.e., derivative with respect to t , not multiplication by t). Then the equation would make sense. Please note that it took me more than 30 minutes to figure this out! This is what happens when authors use inappropriate notations or auto-convert formulas without proofreading them!

(5) Please motivate the left-most inequality of Assumption 1 in more detail. A natural assumption would be $c_a < c_e$ instead of $c_a/(1-\lambda) < c_e$. I am aware that the authors require the scaling by $(1-\lambda)$ to conduct their proofs. Nevertheless, the assumption has to be motivated from the perspective of the application under consideration. If this is a technical assumption it should be stated as such.

(6) Along the same lines as (5): Which of the presented theorems still hold if $c_a < c_e$ but $c_a/(1-\lambda) \geq c_e$? What is the interpretation of such a scenario?

(7) "Due to Assumption 2, both $C_D(t_D)$ and $C_A(t_A)$ are in the form given in (4)." I think this statement is not entirely correct. For C_A to be in the form given in (4) you require that $c_f > c_l$. This, however, follows from Assumption 1, not from Assumption 2.

(8) In the last line of p. 10 I would expect $(1-\lambda)c_f$ instead of $(1-\lambda)c_e$, otherwise the result is not consistent with eq (8).

(9) "The second part of the lemma follows from Lemma 2". To me it is not clear which part of which (of all the unnumbered lemmas) is meant by "second part of the lemma". Please clearly structure the Lemmas/Propositions you formulate, e.g. in the following form:

"Lemma X:

Assume

Then

(i)

(ii)

"

Use (i), (ii) , ... in the formulations instead of referring to "first part",

"second part"

(10) The formulation of Claim 1 is imprecise. What is meant by "otherwise"? When looking at the proofs I think by "otherwise" the authors mean that

$c_e - c_a < P_E(TP)\lambda[c_l - c_f] \dots < c_e$, but this should be explicitly said.

(11) I think in the expression for $C^*D|I=\infty$ (p. 11) " $\lambda c_e + \lambda c_e$ " has to be replaced by " $\lambda c_e + \lambda c_l$ ".

(12) I think in eq. (12) the authors use assumption 2 without stating it.

(13) In mathematics "iff" stands for "if and only if". On p. 12 the authors claim that " $t_1 < 1$ iff $P_E \dots < c_e$ ", however they only prove that $P_E \dots < c_e$ implies $t_1 < 1$, the proof of the second statement ($t_1 < 1$ implies $P_E \dots < c_e$) is missing. Or do the authors mean " $t_1 < 1$ if $P_E \dots < c_e$ " instead of " $t_1 < 1$ iff $P_E \dots < c_e$ "?

(14) On p.12 "for all I" has to be replaced by "for all I from $[0, \infty]$ "

(15) The author claim "First, if $P_E(TP) \dots < c_e - c_a$, then $t_1 < 1$ " (p. 12). Please provide all details necessary to conclude that it holds $t_1 < 1$.

(16) The authors claim "Next, we assume $c_e - c_a < \dots < c_e$, then $t_1 < 1$ " Please provide all details necessary to conclude that it holds $t_1 < 1$.

(17) "This in turn implies $C_E < C_D^*|I=0$ ". This statement is clearly WRONG. The author have shown that " $C_D^*|I=0 - C_E \geq 0$ ", this implies $C_E \leq C_D^*|I=0$ but NOT $C_E < C_D^*|I=0$. Such inaccuracies are typical for high school students. I am very surprised to find them in a submission to an international journal.

(18) A related issue reappears on p. 13. The authors claim " $C_D^*|I=\infty < C_E < C_D^*|I=0$ ", however, what they show is $C_D^*|I=\infty \leq C_E \leq C_D^*|I=0$. At this point I ask myself once more, whether any of the authors has carefully proofread this supplement. According to my understanding scientific peer-reviewing is not intended to be cost-free proofreading.

(19) On p.13 "for all I" has to be replaced by "for all I from $[0, \infty]$ "

(20) p. 14 "Similarly, we have ..." Please provide the reasoning in full detail.

(21) p. 14 "We can easily show that ..." Please provide the reasoning in full detail.

(22) The same issue as raised in (17) and (18) reappears on p. 14.

(23) I kindly request the authors to recapitulate the difference between "decreasing" and "strictly decreasing" functions. See e.g., https://en.wikipedia.org/wiki/Monotonic_function

The statement "the fact that C^*_A is decreasing in I implies it exists a unique I_{ED} " is MISLEADING. For I_{ED} to be unique C^*_A has to be a STRICTLY decreasing function! Actually, on p. 10 the authors show that C^*_A is strictly decreasing

(derivative < 0), therefore, the conclusion is correct, however the argument is phrased incorrectly. Please replace the statement by "the fact that C^*_A is strictly decreasing in l implies ...". Again, I am surprised to have to discuss such basic issues here!

(24) Please provide a detailed calculation of $D'(l)$.

(25) "... then $D'(0) < \text{infinity}$, $D'(\text{infinity}) > 0$ which implies that $D(l)$ is always increasing". This statement is CLEARLY WRONG! Set, e.g., $D=(x-1)^2$. Then $D'(0)=-1 < \text{infinity}$ and $D'(\text{infinity})=\text{infinity} > 0$, but D is not always increasing.

(26) " $D'(0) < -\text{infinity}$ " this is CLEARLY WRONG. A real number cannot be smaller than $-\text{infinity}$.

(27) "Hence, for the low liability case, a necessary and sufficient condition for the existence of l_{DA} " Please provide a detailed reasoning! Does this statement refer only to the case $t_2 < 1$, if yes state explicitly.

(28) "Equation (22) implies" Please provide a detailed step-by-step argumentation supporting the given conclusions.

(29) "If for $l = l_{DA}$, we have $\Delta DA(l)_{cl} < 0$, then around a neighborhood of cl we have $\Delta DA(l_{DA}) < 0$ and $\Delta DA(l_{DA}) < 0$. Since $\Delta DA(l)$ is either always increasing or first decreasing and $\Delta' DA(l_{DA}) > 0$ then l_{DA} should increase with cl ." This statement makes no sense to me. Please restructure it and provide a clean step-by-step reasoning. What is Δ' ?

(30) "Hence, as we increase c_a both C^*_D and C^*_A will increase and C_E will remain constant. Therefore, (23) will continue to hold." The provided argument is INVALID. If C^*_D increases quicker than C^*_A then it could happen that for some values of c_a it holds $C^*_D > C^*_A$. Please provide a valid argument to support your claim.

(31) "The result then follows from a similar argument as in Part 2" Please provide a detailed argumentation supporting your claim.

(32) "Next note that $\lim \dots$ " (p. 22) Please provide a detailed explanation.

(33) "To prove the first part of the lemma". To me it is unclear to what "lemma" the authors refer here. I skip the remainder of the proof of Proposition 2.

Minor:

(a) The notation " $c_s = -c_s > 0$ " is mathematically wrong! If $c_s = -c_s$ then it follows that $c_s = 0$. Please write instead "We substitute c_s by $-c_s$. Then it holds $c_s > 0$ "

(b) For the sake of consistency C_E^* in eq (11) should be replaced by C_E .

(c) p. 15 upper half: please typeset " R^+ " correctly.

(d) there are multiple typos in the first half of p. 22

(Remarks on code availability)

Since I have been asked by the editorial team to focus on the model and proofs in the appendix, I have not reviewed the code.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I am satisfied by the authors' comments in response to my initial feedback.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

Thank you for addressing my comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

This is a well written paper and I feel the authors have addressed the comments raised by the reviewers

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I thank the authors for implementing the suggested changes. The revision has improved the manuscript significantly. The new version of the appendix is much easier to follow. I only have the following MINOR comments:

p. 1. of the supplement (Table 1): I think "False negative (FN) / false positive (FN) /true positive (FN) /true negative (FN)" has to be replaced by "False negative (FN) / false positive (FP) / true positive (TP) / true negative (TN)"

p. 10 of the supplement: "Finally, by letting $t = t^*$ in (6), we can calculate ..."

I find this statement difficult to follow, since (6) does not depend on the variable t . Please adjust the statement or explain it in more detail.

p. 17 of the supplement, second line of equation (33): I think " $+ P_E(TP)\lambda[c_f - c_l]$ " has to be replaced by " $+ P_E(TP)\lambda[c_l - c_f]$ "

p. 24 of the supplement: I think " $D(l) > 00$ " should read " $D(l) > 0$ ".

(Remarks on code availability)

I had a very quick look at the code. The code provides a README file. The code is reasonably commented. I have not checked the details of the implementation. Since the authors cannot share the original data and did not upload the simulated data (which is available upon request, " However, upon request we can share a simulated data that is shaped similarly to the dream data used in the paper.") I have not tried to run the code on my machine.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <https://creativecommons.org/licenses/by/4.0/>

Response Letter

Response to Reviewer #1

*Reviewer #1 (Remarks to the Author): Expert in AI in healthcare and health economics
Thank you for the opportunity to review this manuscript, which seeks to provide generalizable rules for selecting breast cancer screening strategies using AI and/or radiologist expertise. The paper is generally clearly written. I especially appreciated the exploration of asymmetrical liability, for which the authors provided a nice motivation. The use of DREAM Challenge data was also helpful, and I appreciated the author's point that their work might allow for better personalization of strategies based on risk.*

Thank you for your thoughtful and constructive feedback on our manuscript. We are pleased that you appreciated our work and found the context, particularly the liability analysis, interesting.

There are a few points where readers might benefit from more explanation or a more thorough exploration of assumptions.

1) Material and Methods / Analytic Model: You write: "...the delegation strategy in which humans and machines collaborate in decision-making." This doesn't seem like an accurate description of the delegation strategy since the radiologist's determination is (at least theoretically) independent of the AI's.

Thank you for the feedback. You are correct. Our proposal focuses on delegation, in which the AI algorithm performs an initial screening of all cases and then delegates certain ones to humans for final decision-making. We have removed the use of the terms "collaboration" and "collaborate," except in one section where we discuss the broader potential for humans and machines to work together to complete tasks. Instead, we use human-machine combined solutions when needed. Related to this comment, we also changed the title to make it consistent with the revised presentation in the manuscript.

2) Please provide more information about your assumptions on litigation cost. In the supplementary material, you seem to multiply litigation cost by the prevalence of disease, which I don't understand

Thank you for highlighting this important point. For every case that is screened, we wish to compute the expected litigation costs. This consists of the following three components:

1. The average cost that is incurred by the hospital if the client chooses to file a lawsuit
2. The probability that the client will file a lawsuit
3. The probability that the hospital will incorrectly label a cancer patient as healthy (the probability of a missed cancer diagnosis.) This probability is, in turn, the probability

of having the disease (disease prevalence) times the probability of missing the disease, given that it exists in the patient.

The expected cost incurred by the hospital if a cancer is missed is, therefore, the product of the first two components and is represented by the variable c_l . But to arrive at the final expected cost of litigation, this product (c_l) must be multiplied by the probability of a missed diagnosis computed as described above.

In addition, we would like to point out that in our analysis, we have chosen to only consider litigation costs when a cancer patient is incorrectly labeled as healthy (and not when a healthy patient undergoes follow-up procedures unnecessarily, a false-positive scenario). This decision is based on the assumption that litigation primarily arises when cancer is missed (i.e., when a cancer case is incorrectly classified as negative). Our literature review on breast imaging malpractice confirms that litigation primarily results from false negatives, where cancer is missed despite its presence. In general radiology practice, positive assessments rarely lead to litigation (Brady et al. 2017). According to Murphy (2018), approximately 93% of litigation cases are due to delays in cancer diagnosis (false negatives), while only 3.5% are due to false positives, and the remaining 3.5% stem from lack of informed consent. False-positive litigation often concerns procedural issues during follow-up rather than the initial screening stage, which further justifies our assumption.

We include the following in the manuscript regarding this point.

“We assume that litigation arises when cancer is missed (i.e., when a cancer case is incorrectly classified as negative). Our literature review on breast imaging malpractice supports this assumption that litigation mostly results from false negatives (93% of cases are due to delays in cancer diagnosis), where cancer is missed despite its presence.”

Brady, A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable? *Insights Into Imaging* 8 (1), 171–182.

Murphy, B. L., M. D. Ray-Zack, P. N. Reddy, A. J. Choudhry, M. D. Zielinski, E. B. Habermann, L. E. Jakub, K. R. Brandt, and J. W. Jakub (2018). Breast cancer litigation in the 21st century. *Annals of Surgical Oncology* 25 (10), 2939–2947.

3) The assumption of equal variances in the mean score of sick and healthy patients seems worthy of more exploration and/or validation. Does the DREAM Challenge data allow you to validate that assumption? If it does and the assumption proves unreasonable, I'd encourage you to substitute a more data-driven parameter for the variances or to find a way of estimating how sensitive your results are to differences in variance.

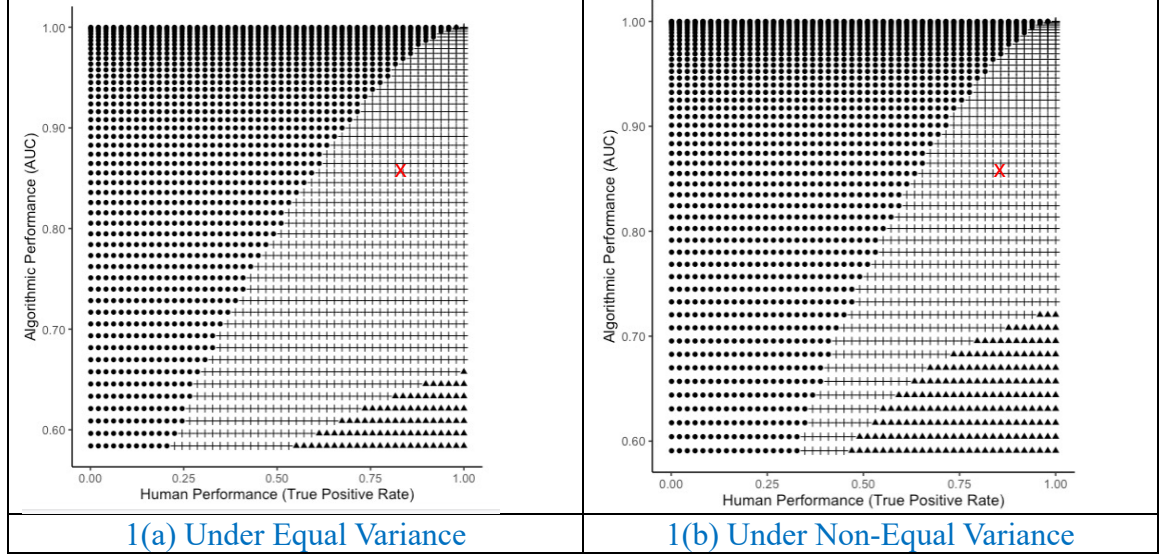
Following your recommendation, we analyzed the DREAM Challenge data and found that the variances of the mean risk scores for cancer and healthy patients were statistically similar for eight out of eighteen algorithms ($p\text{-value} > 0.05$) but significantly different for the others ($p\text{-value} < 0.05$) using the F-test. Therefore, the equal variance assumption is justified in some cases, supporting the feasibility of the analytical model. Aside from being feasible, the equal variance assumption simplifies the mathematical analysis.

In the empirical section, we conduct two types of simulations. First, we use the closed-form solution under the equal variance assumption to identify the optimal policy, as shown in Figure 1(a) below. To explore the impact of unequal variances, we assume that the standard deviation for healthy patients is 0.84 times that of cancer patients (the median proportion across algorithms in the DREAM Challenge data) and obtain Figure 1(b) below. Comparing the figures, the qualitative nature of our theoretical results, as in Proposition 1, remains consistent, and the optimal strategy still indicates delegation as the best approach (the red marks on the figure).

The second type of simulation, also included in the paper, does not rely on any assumption of variances. We backtest the optimal policy using the predictions provided in the competition without enforcing equal variances, assuming real data parameters, and conducting an exhaustive search over-optimization parameters across different strategies. Hence, the numerical section provides results that are free of such assumptions.

We acknowledge the above in the manuscript: " The equal variance assumption holds in eight out of eighteen algorithms from the mammography crowdsourcing challenge data we used in the empirical section. If this assumption is relaxed, optimal costs change, but the theoretical insights remain consistent. We also note that the backtesting experiments in the numerical section do not enforce equal variances."

Figure 1. Optimal Policy in The Parameter Space



4) In Table 1a, the optimal strategy is defined by balanced accuracy (BA) and net financial benefit. Could you provide an intuition for how you arrived at the BA cutoff used in the table?

Thank you for bringing this to our attention. Finding the optimal strategy requires minimization across three cost functions as in the following:

$$C^* := \min\{C_E, C_D^*, C_A^*\},$$

where $C_D^* := \min_{t_D} C_D(t_D)$ and $C_A^* := \min_{t_A} C_A(t_A)$. Solving the overall optimization problem requires a comparison of the optimal solution of subproblems and the expected cost of the expert-only strategy, C_E . Equations (24), (25), and (26) in the supplement present the closed-form expressions corresponding to the three components of C^* . The thresholds presented in Table 1a naturally arise from solving the models and the three-way comparisons. We would also like to clarify that we did not explicitly include balanced accuracy (BA) and net financial benefit in the model; these measures directly follow from our derivation and provide intuitive means to characterize the optimal solution across the parameter space. The reason why we have chosen to present the results in terms of the BA and the net financial benefit is because the two are common metrics used by the value of information and medical decision-making literature. Let us elaborate.

Our cost optimization model incorporates statistical principles of machine learning with underlying distributions for sick and healthy patients. This setup is analogous to the **value of information** under uncertainty, a well-established concept in medical decision-making (Felder and Mayrhofer 2017). Our model evaluates the trade-offs around using AI predictions alone vs. delegating the decision to a radiologist for a given mammogram interpretation task. The **balanced accuracy** threshold in our model is linked to the discriminatory power of AI, similar to the Youden index, which is often used to measure

diagnostic performance. The net financial benefit is analogous to the net benefit quantified in decision curve analysis (a method evaluating the **value of predictive information**), reflecting the trade-off between the costs and benefits of implementing the AI-based strategy (Vickers and Elkin 2007).

Rearranging our results within the context of balanced accuracy and net financial benefit provides an intuitive balance between accuracy and financial impact and represents a novel characterization of AI use for decision making in the context of the delegation strategy. To our knowledge, this specific application of AI and the resulting trade-offs have not been previously explored in the literature.

Felder, Stefan, and Thomas Mayrhofer. Medical decision making. Springer Berlin Heidelberg, 2017.

Vickers, Andrew J., and Elena B. Elkin. "Decision curve analysis: a novel method for evaluating prediction models." *Medical Decision Making* 26, no. 6 (2006): 565-574.

To prevent confusion, we include the following in the revised manuscript.

“Drawing from the value of information and medical decision-making literature [8,9], the characterization of optimal strategy in the parameter space results in intuitive thresholds also seen in these streams of literature. Hence, for the clarity of presentation, we define the following terms naturally derived from our model.”

5) *Typo in Table 2 ("Medicate" instead of "Medicare").*

Thank you for identifying this typo. We have now corrected it.

6) *Does your work rely on an assumption that AI and radiologists would detect the same set of cases? It seems like a delegation strategy might be cost-minimizing for a low sensitivity AI if it were to detect only the cases that are most likely to be missed by radiologists, for example. Making that assumption explicit (or correcting my intuition) would be helpful.*

Thank you for your feedback. First, we would like to note that the AI algorithm provides a continuous risk assessment. Hence, the sensitivity and specificity of the algorithm would be risk threshold dependent. The sensitivity and specificity for the radiologist, on the other hand are drawn from population data. Second, our analytical model is based on the assumption of conditional independence between the AI and radiologists given cancer status. This means that, given the cancer status, the predictions of the AI and radiologists are independent of each other (implicitly, this means the radiologist and AI will make independent mistakes and not necessarily detect the same cases). Consequently, we do not assume that the AI or radiologists have superior sensitivity in detecting cancer. Instead, our assumption reflects the potential for diverse diagnostic capabilities between AI and radiologists. Finally, backtesting experiments in the numerical section uses AI predictions provided in the competition without enforcing the assumptions of the analytical model and

conducting an exhaustive search over optimization parameters across different strategies. Hence, the numerical section provides results that are free of such assumptions. We have now explicitly stated that in the numerical analysis section the following:

“Note that the backtesting experiments do not rely on the assumptions presented in the modeling section.”

Response to Reviewer #2

Reviewer #2 (Remarks to the Author): Expert in AI and machine learning models and mammography analysis

The paper presents the economics of incorporating AI fully or partially into breast screening, primarily considering a US healthcare environment where single reading and litigation are more prominent compared to other geographic jurisdictions. This is an important topic when it comes to integration and deploying AI. The paper's significance lies in the design of the economic model presented and providing insight into the parameters that are important for healthcare institutions to consider when applying the model to their situation to understand how to optimally integrate AI into their system in a way that will work for them and meet their needs.

We thank you and the rest of the review team for the constructive and thoughtful feedback, which helped us significantly improve our paper. We have carefully considered addressing all your comments and questions. A summary of the key changes is included on the first page of this response letter. Below, we provide point-by-point responses to your comments with the original comments in italics and our responses in blue font and indented.

The model designed is presented very well and organised in a way that is easy to follow. It would be worth noting how the USA healthcare environment has steered the design of the model and parameters used, and what adjustments should be considered for other geographies. For example, the strategy where AI replaces the human expert could occur in geographies outside of the USA, however, the strategy where AI replaces only 1 of 2 human experts in geographies where double reading is standard is not considered.

Thank you for this important comment. As you noted, our model is tailored to the U.S. healthcare environment, where single reading is standard, and litigation costs are significant. We recognize that needs to be explicitly stated. Accordingly, we include the following in the manuscript.

“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.”

It’s important to also note that use of AI performance data from the DREAM challenge may now be outdated with the top algorithm demonstrating a 0.857 AUC. Current day AI algorithms in screening mammography can reach AUCs around 0.95. Some medical device intended use indications even require a minimum of a 0.95 AUC for regulatory clearance in the USA. A higher AUC will have an impact on reducing false positives and false negatives and the level of automatic possible with AI. It would be important and interesting to consider the impact of this higher AI performance bound on the economic model.

We agree that AI algorithms are continually improving, with recent advancements surpassing the performance reported in the DREAM Challenge. We have addressed this by adding **a new corollary (Corollary 1)** to our analysis that examines the impact of higher AUC values on our model. This addition confirms your intuition that increased AUC values can significantly reduce false positives and false negatives and enhance the feasibility of automation with AI.

We would also like to note that to see the effects of the evolving AI Performance, our empirical section has two separate parts. In the first part, where we use simulated data, we vary the performance of the AI to investigate various scenarios. For example, our empirical results further support your observation, showing that as AUC improves, the potential for automation increases (see Figure 1). Specifically, with the current AUC of 0.95 for breast screening (de Vries et al. 2023) and a radiologist sensitivity benchmark of 87.4% (Lee et al. 2023), the optimal strategy shifts to automation when algorithm costs are low; otherwise, delegation remains preferable. We highlight this point in the empirical section of our manuscript.

Lee CI, Abraham L, Miglioretti DL, Onega T, Kerlikowske K, Lee JM, Sprague BL, Tosteson ANA, Rauscher GH, Bowles EJA, diFlorio-Alexander RM, Henderson LM; Breast Cancer Surveillance Consortium. National Performance Benchmarks for Screening Digital Breast Tomosynthesis: Update from the Breast Cancer Surveillance Consortium. Radiology. 2023 May;307(4):e222499. doi: 10.1148/radiol.222499. Epub 2023 Apr 11. PMID: 37039687; PMCID: PMC10323294. [[Abstract and Full Text](#)]

Fluss, Ronen, David Faraggi, and Benjamin Reiser. "Estimation of the Youden Index and its associated cutoff point." *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 47.4 (2005): 458-472.

de Vries, C.F., Colosimo, S.J., Staff, R.T., Dymiter, J.A., Yearsley, J., Dinneen, D., Boyle, M., Harrison, D.J., Anderson, L.A., Lip, G. and iCAIRD Radiology Collaboration, 2023. Impact of different mammography systems on artificial intelligence performance in breast cancer screening. *Radiology: Artificial Intelligence*, 5(3), p.e220146.

Response to Reviewer #3

Reviewer #3 (Remarks to the Author): Expert in mammography, breast cancer radiology, and AI

I enjoyed reading this well written paper but I have some questions

We thank you and the rest of the review team for the constructive and thoughtful feedback, which helped us significantly improve our paper. We have carefully considered addressing all your comments and questions. A summary of the key changes is included on the first page of this response letter. Below, we provide point-by-point responses to your comments with the original comments in italics and our responses in blue font and indented.

1. What model of film reading did the authors use - single or double reading

Thank you for bringing up this point. We added the following sentence to the manuscript to clarify this point:

“The dataset uses 2D digital screening mammograms interpreted by a single radiologist”

2. Did the authors consider the expertise of the readers as the recall regarding a breast radiologist will be lower than a non breast radiologist

Thank you for your comment. Our theoretical model accommodates various performance levels using two separate parameters for radiologist performance, including those of both breast imaging experts and general community radiologists. A healthcare organization can design its mammography operations based on the average performance of radiologists it employs (can be a mix of breast and non-breast radiologists) or that of a single radiologist (a more personalized approach).

In the second part of our numerical experiments, AI performance from the DREAM Challenge is below the benchmarks of U.S. community radiologists. We recognize that expertise impacts recall rates, and while AI algorithms are improving—current AUCs of 0.95 for breast screening (de Vries et al., 2023) and radiologist sensitivity benchmarks of 87.4% (Lee et al., 2023)—the optimal strategy is automation when algorithm costs are low; otherwise, delegation remains preferable. We address this in the empirical section of our manuscript. Specifically, we write:

“These outcome probabilities reflect the average performance of radiologists, such as the typical performance for a given mix of radiologist expertise within a healthcare organization, or the performance of an individual radiologist if the organization opts for a more personalized approach.”

Lee CI, Abraham L, Miglioretti DL, Onega T, Kerlikowske K, Lee JM, Sprague BL, Tosteson ANA, Rauscher GH, Bowles EJA, diFlorio-Alexander RM, Henderson LM; Breast Cancer Surveillance Consortium. National Performance Benchmarks for Screening Digital Breast Tomosynthesis: Update from the Breast Cancer Surveillance Consortium. *Radiology*. 2023 May;307(4):e222499. doi: 10.1148/radiol.222499. Epub 2023 Apr 11. PMID: 37039687; PMCID: PMC10323294. [[Abstract and Full Text](#)]

de Vries, C.F., Colosimo, S.J., Staff, R.T., Dymiter, J.A., Yearsley, J., Dinneen, D., Boyle, M., Harrison, D.J., Anderson, L.A., Lip, G. and iCAIRD Radiology Collaboration, 2023. Impact of different mammography systems on artificial intelligence performance in breast cancer screening. *Radiology: Artificial Intelligence*, 5(3), p.e220146.

3. do the authors know the rates of interval cancers - this gives an idea of the cancers missed by human readers and how does this translate for the algorithm as this will impact on litigation costs

Thank you for raising this relevant point. We recognize that accounting for interval cancer rates can help in estimating litigation costs more accurately. Unfortunately, our data does not specify which cases are interval cancers, and we therefore cannot identify interval cancers individually. In our dataset, women who were not recalled but later had clinically detected cancers before their next scheduled mammogram are classified as having false-negative screening exams.

Although the information on interval cancers is not present in our dataset, the theoretical model indirectly handles interval cancers. The variable c_l in our model represents the expected litigation cost due to a missed cancer diagnosis, encompassing various reasons for litigation, including the likelihood of a lawsuit.

“This expected litigation cost accounts for population heterogeneity in key factors, such as the occurrence of interval cancers and variations in cancer grades.”

4. The authors talked about missing cancers, but another caveat is the grade of cancer missed - missing grade 1 versus grade 3 - I am not sure if that can be factored in

This is another relevant point which we appreciate. We acknowledge that the grade of a missed cancer could potentially influence litigation costs, as higher-grade cancers (being more aggressive with potentially worse outcomes), as compared with lower-grade ones, may have higher costs due to perceived delays in diagnosis or mismanagement such cancers. This distinction in cancer grades, in theory, may also relate to different radiologists and AI performance. However, our current model and empirical analysis do not account for cancer grade as we do not have such information in our data.

We would also like to note that our model focuses on the screening stage, where the primary goal is detecting abnormalities, consistent with most AI models available in practice that aim to distinguish normal from abnormal mammograms. Grading and staging occur in downstream tasks, handled by different AI models, and we are not aware of research predicting cancer grades based solely on screening mammograms. Our model considers average heterogeneity in grade and stage to estimate litigation costs at the population level. More granular data on cancer grades could potentially enhance our model and offer a clearer understanding of litigation risks associated with different cancer grades, as well as how AI models can be incorporated into mammography workflows.

5. how did the algorithms assess risk ? was it binary or scale of 0-100 and how was the threshold set

Thank you for your question. Based on your comment, we now clarify in the manuscript that “the algorithms assessed risk using a continuous scale from zero to one, representing the probability of cancer.” This continuous nature of the risk score facilitates our proposed delegation strategy. The threshold for classification is determined through our mathematical model, which minimizes overall costs by optimizing the balance between false positives and false negatives.

Response to Reviewer #4

Reviewer #4 (Remarks to the Author): Expert in mathematical modelling, computational medicine and oncology

We thank you and the rest of the review team for the constructive and thoughtful feedback, which helped us significantly improve our paper. We apologize that a formatting error when switching the latex template led to some confusion, which is now fixed in this revised version. Moreover, we have rewritten most parts of the proofs section to increase their clarity, as you suggested. We have carefully considered addressing all your comments and questions. A summary of the key changes is included on the first page of this response

letter. Below, we provide point-by-point responses to your comments with the original comments in italics and our responses in blue font and indented.

As requested by the editorial team, I focused on the proofs presented in the Appendix. The presented model is straightforward and not very involved. From the mathematical point of view, the proofs are on a level which a talented high school student can accomplish, or, alternatively, everyone who has passed an introductory calculus course. For the topics under investigation this methodology is appropriate.

However, the style of presentation is very poor. I have severe doubts that any of the authors has proofread the appendix carefully. Besides subtle issues in the reasoning (which can be easily corrected) the presentation lacks structure, the notation is inappropriate at important spots and basic terminology is used incorrectly. Arguments are partially wrong or insufficient.

In its current form the appendix is clearly unacceptable for publication and requires major revisions before it can be reevaluated.

We are extremely apologetic about the lack of clarity and the errors in the appendix. We truly appreciate your thorough feedback and have carefully rewritten the proofs in the appendix to improve its clarity. We have made every effort to address each and every one of your concerns. In some cases, we are able to provide a direct response to your specific point. Occasionally, since the revision required complete rewriting, we provide a general response and describe how we clarified to the extent it can be presented in the response letter (while also referring to specific section in the supplement). Thank you again for your time and effort in providing such detailed feedback.

Below I list some of the issues I encountered when reading the supplement. Since the authors have neither included line numbers nor have they numbered all equations, it is challenging to refer to a precise spot in the text.

We have now ensured that every essential equation is numbered and that any reference to an equation is made using the equation number.

(1) Many of the proofs refer to “lemma 1” and “lemma 2”. However, there exist no statements which are labeled as “lemma 1” and “lemma 2”. Of course, I can guess from the context, which statements the authors have in mind, when referring to “lemma 1” and “lemma 2”, but for the sake of readability please label the respective statements proven on pp. 7 as “Lemma 1” and “Lemma 2”.

You are correct; this was an oversight in our initial submission. We inadvertently omitted the labels for the two lemmas and apologize for that omission. We have now labeled the relevant statements as "Lemma 1" and "Lemma 2" in the revised supplement. Thank you for pointing this out.

(2) Please use one of the established standard notations to denote derivatives. It is unacceptable to denote the derivative of $C(t)$ with respect to t as $C(t)t$ (see eq. 7 for example) because this looks as the multiplication of $C(t)$ and t . Use either $C'(t)$ or $dC(t)/dt$. The same applies to the last equation on the bottom of p. 7 and to many other equations (e.g., eq. 10, eq 22, unnumbered equation at the bottom of p. 19, eq 27)

We agree. We have now fixed this issue throughout the paper. We now make sure to use standard notation for every derivative and partial derivative used in the model and its proofs.

(3) The statement of a Proposition or Lemma should mention all assumptions which are required and not only the conclusion. It not acceptable to present assumptions between the lines and use them e.g. ten or more pages later without clearly stating it.

We agree. Following your recommendation, we now explicitly state Assumptions 1 and 2, which are used throughout the analysis, as well as all assumptions required for each Proposition and Lemma in respective statements of those. These assumptions are labeled and referenced with equation numbers in the proofs, enhancing clarity and accessibility.

(4) The third equation on p.8, i.e., the equation without number which follows equation (7) is clearly incorrect. For $t=0$ the left hand-side equals 0 whereas the right hand-side is different from zero. I guess instead of “(.....) t ” the authors mean “ $d (...) / dt$ ” (i.e., derivative with respect to t , not multiplication by t). Then the equation would make sense. Please note that it took me more than 30 minutes to figure this out!

This is what happens when authors use inappropriate notations or auto-convert formulas without proofreading them!

We apologize for the inconvenience and appreciate your effort in identifying the error. We have corrected this and thoroughly reviewed the entire supplement to prevent similar issues.

(5) Please motivate the left-most inequality of Assumption 1 in more detail. A natural assumption would be $c_a < c_e$ instead of $c_a / (1 - \lambda) < c_e$. I am aware that the authors require the scaling by $(1 - \lambda)$ to conduct their proofs. Nevertheless, the assumption has to be motivated from the perspective of the application under consideration. If this is a technical assumption it should be stated as such.

We have clarified that the inequalities in Assumptions 1 and 2 are technical assumptions necessary for our proofs. Specifically, we state: “In solving our optimization model, we apply two technical simplifying assumptions, which, while technical, are practically satisfied, as detailed below.” Regarding the example, these assumptions hold in breast cancer screening application due to low disease prevalence (e.g., breast cancer, where $\lambda = 0.00128$).

(7) “Due to Assumption 2, both $C_D(t_D)$ and $C_A(t_A)$ are in the form given in (4).” I think this statement is not entirely correct. For C_A to be in the form given in (4) you require that $c_f > c_l$. This, however, follows from Assumption 1, not from Assumption 2.

Thank you for pointing this out. We have corrected the statement to reference Assumption 1, as the conclusion $C_D(t_D)$ and $C_A(t_A)$ are in the form given in (4) follows from Assumption 1, not Assumption 2.

(8) In the last line of p. 10 I would expect $(1-\lambda)c_f$ instead of $(1-\lambda)c_e$, otherwise the result is not consistent with eq (8).

You are correct. We now fixed this typo.

(9) “The second part of the lemma follows from Lemma 2”. To me it is not clear which part of which (of all the unnumbered lemmas) is meant by “second part of the lemma”. Please clearly structure the Lemmas/Propositions you formulate, e.g. in the following form:

“Lemma X:

Assume

Then

(i)

(ii)

“

Use (i), (ii) , ... in the formulations instead of referring to “first part”, “second part”

Thank you for your feedback. We have revised the structure of our Lemmas and Propositions to use clear labels (e.g., (1), (2)) instead of referring to "first part" and "second part," as you suggested. This should clarify which specific results we are referring to. We then refer to these labels in the ensuing discussion.

(10) The formulation of Claim 1 is imprecise. What is meant by “otherwise”? When looking at the proofs I think by “otherwise” the authors mean that $c_e - c_a < P_E(TP)\lambda[c_l - c_f] - \dots < c_e$, but this should be explicitly said.

Following your recommendations, we now explicitly state the conditions rather than writing otherwise for all the Claims in that Proposition. For example, for Claim 1, we write:

Otherwise, if

$$c_e - c_a \leq P_E(TP)\lambda[c_l - c_f] - P_E(FP)(1 - \lambda)c_f, \quad (18)$$

then there exists a unique $I_{ED} \in [0, \infty]$ s.t.

$$C_E \leq C_D^*(I), \text{ for } 0 \leq I \leq I_{ED}, \text{ and } C_D^*(I) < C_E, \text{ for } I > I_{ED}. \quad (19)$$

(11) I think in the expression for $C^*D|I=\text{infinity}$ (p. 11) " $\lambda c_e + \lambda c_e$ " has to be replaced by " $\lambda c_e + \lambda c_l$ ".

Thank you for pointing out this typo; we have now fixed it.

(12) I think in eq. (12) the authors use assumption 2 without stating it.

We now add to the text that this inequality follows from assumption 2. We have also changed the text in other locations in the paper to further clarify the exposition and the assumptions required in the proof.

(13) In mathematics "iff" stands for "if and only if". On p. 12 the authors claim that " $t_l < 1$ iff $P_E \dots < c_e$ ", however they only prove that $P_E \dots < c_e$ implies $t_l < 1$, the proof of the second statement ($t_l < 1$ implies $P_E \dots < c_e$) is missing. Or do the authors mean " $t_l < 1$ if $P_E \dots < c_e$ " instead of " $t_l < 1$ iff $P_E \dots < c_e$ "?

Thank you for this constructive comment. Based on your feedback, we have rewritten the proof of this theorem. The refined theorem does not require this if and only if condition.

(14) On p.12 "for all I" has to be replaced by "for all I from $[0, \text{infinity}]$ "

Thanks for the suggestion. We have fixed this based on your suggestion.

(15) The author claim "First, if $P_E(TP) \dots < c_e - c_a$, then $t_l < 1$ " (p. 12). Please provide all details necessary to conclude that it holds $t_l < 1$.

Based on this suggestion, we have redefined various parameters such as t_1 and t_2 throughout the text and used those definitions to provide all the necessary steps to conclude our inequalities. We present the details in the Equation set (24) on page 15 of the Supplement.

(16) The authors claim "Next, we assume $c_e - c_a < \dots < c_e$, then $t_l < 1$ " Please provide all details necessary to conclude that it holds $t_l < 1$.

Thank you very much for pointing this out. Equation set (24) on page 15 of the Supplement shows the details for the claim to hold.

(17) "This in turn implies $C_E < C_D^*|I=0$ ". This statement is clearly WRONG. The author have shown that " $C_D^*|I=0 - C_E > 0$ ", this implies $C_E \leq C_D^*|I=0$ but NOT $C_E < C_D^*|I=0$. Such inaccuracies are typical for high school students. I am very surprised to find them in a submission to an international journal.

Thank you for pointing out this typo. We have revised the entire supplement to correct such inaccuracies and ensure the quality of the manuscript. We assure you that we have now gone through the proofs with great care.

(18) A related issue reappears on p. 13. The authors claim " $C_D^*|I=\infty < C_E < C_D^*|I=0$ ", however, what they show is $C_D^*|I=\infty \leq C_E \leq C_D^*|I=0$. At this point I ask myself once more, whether any of the authors has carefully proofread this supplement. According to my understanding scientific peer-reviewing is not intended to be cost-free proofreading.

We apologize once again for this mishap. We have rewritten and proofread the whole supplement. For your specific comment, we have now changed the inequalities to make them mathematically correct, as you suggested.

(19) On p.13 "for all I" has to be replaced by "for all I from [0,infinity]"

We have corrected this at various places in the text.

(20) p. 14 "Similarly, we have ..." Please provide the reasoning in full detail.

Based on your feedback, we have rewritten the proofs to include detailed derivations of the results. The specifics addressing your points are now provided in equations (40), (41), and (42) on page 19 of the revised Supplement. We hope you find the updated proofs satisfactory.

(21) p. 14 "We can easily show that ..." Please provide the reasoning in full detail.

Thanks again. We have now provided a detailed reasoning for this in equations (44) and (45) on page 20 of the revised Supplement.

(22) The same issue as raised in (17) and (18) reappears on p. 14.

Thanks for bringing this issue. In the revised version, we have now fixed the problems with inequalities.

(23) I kindly request the authors to recapitulate the difference between "decreasing" and "strictly decreasing" functions. See e.g., https://en.wikipedia.org/wiki/Monotonic_function. The statement "the fact that C_A^* is decreasing in I implies it exists a unique I_{ED} " is MISLEADING. For I_{ED} to be unique C_A^* has to be a STRICTLY decreasing function! Actually, on p. 10 the authors show that C_A^* is strictly decreasing (derivative <0), therefore, the conclusion is correct, however the argument is phrased incorrectly. Please replace the statement by "the fact that C_A^* is strictly decreasing in I implies ...". Again, I am surprised to have to discuss such basic issues here!

We appreciate your feedback on this issue and agree that this needs to be clarified. We now fixed these issues as they appear in the manuscript and improve the mathematical precision of the proofs. We use the term strictly increasing or strictly decreasing; we use the terms

non-increasing or non-decreasing when referring to weekly decreasing and weakly increasing, respectively. We indicated this usage at the beginning of the supplement to make it clear. Specifically, we write:

“When presenting our results, we use “increasing” (or “decreasing”) in the strict sense, and “nondecreasing” (or “nonincreasing”) to indicate weakly increasing or decreasing.”

(24) Please provide a detailed calculation of $D'(I)$.

Thank you for your comment. We include the detailed calculation in the manuscript as below:

from (48), we have $D(\infty) > 0$. Let K and $G(I)$ be defined as

$$K := \frac{e^{-\frac{\ln^2(t_2)}{2I^2} - \frac{I^2}{8}}}{\sqrt{2\pi}} \times \sqrt{\lambda c_f(1-\lambda)(c_l - c_f)}$$

$$G(I) := \left(1 - \underbrace{\frac{\sqrt{P_E(TP)[c_l - c_f] - c_e}}{\sqrt{c_l - c_f}}}_{< 1 \text{ since } P_E(TP) < 1} \underbrace{\frac{\sqrt{c_e + P_E(FP)c_f}}{\sqrt{c_f}}}_{< 1 \text{ due to Assumption 2}} e^{\frac{\ln^2(t_2) - \ln^2(t_1)}{2I^2}} \right).$$

The derivative of $D(I)$ with respect to I is

$$D'(I) = \underbrace{K}_{> 0} \times \left(1 - \underbrace{\frac{\sqrt{P_E(TP)[c_l - c_f] - c_e}}{\sqrt{c_l - c_f}}}_{< 1 \text{ since } P_E(TP) < 1} \underbrace{\frac{\sqrt{c_e + P_E(FP)c_f}}{\sqrt{c_f}}}_{< 1 \text{ due to Assumption 2}} e^{\frac{\ln^2(t_2) - \ln^2(t_1)}{2I^2}} \right).$$

$$= KG(I) \quad (49)$$

(25) “... then $D'(0) < \text{infinity}$, $D'(\text{infinity}) > 0$ which implies that $D(I)$ is always increasing”. This statement is CLEARLY WRONG! Set, e.g., $D=(x-1)^2$. Then $D'(0)=-10$, but D is not always increasing.

Again, we apologize. The statement should have read $D'(0) > 0$ instead of $D'(0) < \text{infinity}$. Given this correction, the claim that follows is correct.

(26) “ $D'(0) < -\text{infinity}$ ” this is CLEARLY WRONG. A real number cannot be smaller than -infinity.

Thanks for this bringing this typo. This should have been $\lim_{I \rightarrow 0} D'(I) \rightarrow -\infty$. In the revised manuscript, we have now added some more details and intermediate steps, to improve the clarity of the proof.

(27) “Hence, for the low liability case, a necessary and sufficient condition for the existence of I_{DA} ” Please provide a detailed reasoning! Does this statement refer only to the case $t_2 < 1$? If yes, state explicitly?

Thank you for this comment on clarity. We now clearly state conditions on t_2 , and make them consistent with the manuscript. Specifically, we write

“Note that $t_2 < 1 \Leftrightarrow \lambda < \frac{c_f}{c_l}$; we refer to this condition as low-prevalence case.

Similarly, $t_2 \geq 1 \Leftrightarrow \lambda \geq \frac{c_f}{c_l}$ which we refer to as high-prevalence case.

(28) “Equation (22) implies” Please provide a detailed step-by-step argumentation supporting the given conclusions.

We apologize that there was a typo here regarding the partial derivative. This inequality follows from the fact that the erfc function is always positive. We have added this detail in the text.

(29) “If for $I = IDA$, we have $\Delta DA(I)c_l < 0$, then around a neighborhood of c_l we have $\Delta DA(IDA) < 0$ and $\Delta DA(IDA) < 0$. Since $\Delta DA(I)$ is either always increasing or first decreasing and $\Delta' DA(IAD) > 0$ then IDA should increase with c_l .” This statement makes no sense to me. Please restructure it and provide a clean step-by-step reasoning. What is Δ' ?

We now clarified this proof and include it below for your convenience.

Assume for a set of base parameters, we have

$$0.5\Phi\left(0.5I + \frac{\ln(t_2)}{I}\right) < P_E(TP)\Phi\left(0.5I + \frac{\ln(t_1)}{I}\right),$$

as such $\frac{\partial \Delta_{DA}(I, c_l)}{\partial c_l} < 0$. By definition of the derivative, there exists $\epsilon > 0$ such that whenever $c'_l \in (c_l, c_l + \epsilon]$, we have $\Delta_{DA}(I_{DA}, c'_l) < 0$. This implies that whenever $c'_l \in (c_l, c_l + \epsilon]$, we have $C_D^*(I_{DA}) < C_A^*(I_{DA})$. In the proof of Claim 3 of Proposition 1, we demonstrated that for a given c_l , $\Delta_{DA}(I, c_l)$ is either always increasing in I or first decreasing and then increasing in I and $\Delta_{DA}(\infty, c_l) > 0$. Hence, since $\Delta_{DA}(I_{DA}, c'_l) < 0$ when $c'_l \in (c_l, c_l + \epsilon]$ and $\Delta_{DA}(\infty, c'_l) > 0$, there exists a value $I'_{DA} > I_{DA}$ such that $\Delta_{DA}(I'_{DA}, c'_l) = 0$. A similar argument holds for the case when $\frac{\partial \Delta_{DA}(I, c_l)}{\partial c_l} > 0$. $\square$

(30) “Hence, as we increase c_a both C^*_D and C^*_A will increase and C_E will remain

constant. Therefore, (23) will continue to hold.” The provided argument is INVALID. If C^*_D increases quicker than C^*_A then it could happen that for some values of c_a it holds $C^*_D > C^*_A$. Please provide a valid argument to support your claim.

You are correct in that it was not clear. We added to the manuscript stating the following to address this point:

“The results follows from the fact that $C^*_D(I) - C^*_A(I)$ is independent of c_a .

(31) “The result then follows from a similar argument as in Part 2 ” Please provide a detailed argumentation supporting your claim.

We have now written a step-by-step argument for this. For convenience we included the proof below:

Step 3: Assume that $I'' < I$. We know that I'' is independent of c_a , so we have

$$C^*_A(I) < C^*_D(I), \quad \forall c_a > 0.$$

However, $C^*_A(I)$ is increasing in c_a and $C^*_A(I)$ approaches infinity as c_a goes to infinity, whereas C_E remains unchanged. Therefore, there exists a c''_a such that if $c_a < c''_a$, we have

$$C^*_A(I) < C_E,$$

and if $c_a > c''_a$, we have

$$C_E < C^*_A(I),$$

which completes the proof of the corollary. $\square$

(32) “Next note that $\lim \dots$ ” (p. 22) Please provide a detailed explanation.

This follows from the fact that under delegation when the liability cost due to the algorithm approaches ∞ , all cases will be delegated to the human expert. We have now added this reasoning to the manuscript (see page 34).

(33) “To prove the first part of the lemma”. To me it is unclear to what “lemma” the authors refer here. I skip the remainder of the proof of Proposition 2.

This was an unfortunate typo as we refer to Proposition 2, which is now fixed.

Minor:

(a) The notation “ $c_s = -c_s > 0$ ” is mathematically wrong! If $c_s = -c_s$ then it follows that

$c_s=0$. Please write instead “We substitute c_s by $-c_s$. Then it holds $c_s>0$ ”

We have changed this using your suggestion.

(b) For the sake of consistency C_E^* in eq (11) should be replaced by C_E .

Thanks, we have fixed this now.

(c) p. 15 upper half: please typeset “ R^+ ” correctly.

Thanks, we have fixed this now.

(d) there are multiple typos in the first half of p. 22

Thanks, we have fixed those typos now.

Response Letter

Response to Reviewer #1

I am satisfied by the authors' comments in response to my initial feedback.

Thank you for your time in this and the previous round; we are glad our changes addressed your concerns.

Response to Reviewer #2

Reviewer #2 (Remarks to the Author): Thank you for addressing my comments.

Thank you for your time in this and the previous round; we are glad our changes addressed your concerns.

Response to Reviewer #3

Reviewer #3 (Remarks to the Author): This is a well written paper and I feel the authors have addressed the comments raised by the reviewers.

Thank you for your positive feedback. We are grateful for your thorough review and support of our work.

Response to Reviewer #4

Reviewer #4 (Remarks to the Author): Expert in mathematical modelling, computational medicine and oncology

I thank the authors for implementing the suggested changes. The revision has improved the manuscript significantly. The new version of the appendix is much easier to follow. I only have the following MINOR comments:

Thank you for your thoughtful review and for acknowledging the improvements in the revised manuscript and appendix. We are glad to hear that the changes have enhanced the clarity and quality of the work. We appreciate your feedback in the previous round as it

helped us improve the paper significantly. Below please find our comments to the minor comments you brought up.

p. 1. of the supplement (Table 1): I think "False negative (FN) / false positive (FN) / true positive (FN) / true negative (FN)" has to be replaced by "False negative (FN) / false positive (FP) / true positive (TP) / true negative (TN)"

Thank you for catching this error. We have corrected it in the supplement on page 1, Table 1. The text now reads "False negative (FN) / false positive (FP) / true positive (TP) / true negative (TN)" as intended. We appreciate your careful review!

p. 10 of the supplement: "Finally, by letting $t = t^*$ in (6), we can calculate ..."
I find this statement difficult to follow, since (6) does not depend on the variable t . Please adjust the statement or explain it in more detail.

Thank you for pointing this out. You are correct that the reference to equation (6) was incorrect. This has been updated to refer to the correct equation (equation (5)), which does depend on the variable t . We appreciate your careful review.

p. 17 of the supplement, second line of equation (33): I think " $+ P_E(TP)\lambda[c_f - c_l]$ " has to be replaced by " $+ P_E(TP)\lambda[c_l - c_f]$ "

Thank you for pointing out this issue. You are correct that there was a sign error. We have now corrected it as per your suggestion.

p. 24 of the supplement: I think " $D(I) > 00$ " should read " $D(I) > 0$ ".

Thanks for catching this typo. We removed the extra 0.