Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-22-71

Reviewer A

The covered topic and the results are valuable. My comments are:

Comment 1: Line 117: Authors should clarify in the manuscript what is meant by "lack of inclusivity and engagement of the end-users (10)." Does it mean interpretability and explainability of the decisions/predictions produced by AI-driven tools or inclusion of the domain knowledge in the algorithm development phase?

Reply 1: We agree with reviewer A that this sentence needs further clarification. We meant to communicate the second explanation provided by reviewer A and consequently revised the sentence to: *"It has been suggested that a key factor of the poor implementation of AI algorithms is the lack of inclusivity and engagement of the end-users and their domain knowledge during the development of these tools"*

Comment 2: 241-242: "conduct a rigorous study and publish in [high impact journal] that will be adopted worldwide." -meaning of rigorous is ambiguous in this context and may vary from clinician to clinician in definition unless clearly defined and explained what is met by that. Should be specified what ''rigorous" study should be reporting.

Reply 2: we understand reviewers A's comment and agree that the meaning of 'rigorous' can be interpreted in various ways, depending on the clinician. However, this is a quote from one of the interviewees. Providing quotes is standard practice in analyzing and reporting qualitative data. Quotes cannot be paraphrased or changed in other ways: they have to be a direct and exact reflection of what was said by the interviewee. In other words: it has to be copy paste from the transcript. Therefore, to adhere to reporting guidelines and scientific rules, we need to retain the wording as it was said by the interviewee, which was 'rigorous'. We also believe that this comment, and the potential differences in interpretation of 'rigorous' is a fitting reflection of what we explain in that paragraph: *Physicians found evidence strength an important facilitator, although they did not agree on the type 239 of evidence that would be sufficient.* Different physicians need different types of evidence for them to trust an algorithm.

Comment 3: 295-196. "I want to see an algorithm work in practice. To be able to work with it, and know exactly how it operates. I need to build trust. (…)" this is the most common reply from clinicians I experience. The analysis lacks investigating more in detail what should be provided (possibly as the output of the algorithm, besides the prediction) that would meet this requirement, eg. Would provide the rational behind the decision algorithm makes (e.g. to visualise the factors that contributed algorithm to make a certain decision) for each individual

prediction solve this problem?

Reply 3: We are glad to read that the results of our study match the experience of reviewer A. In the quote in line 295-296, the interviewee explains that they would like to work with an algorithm and compare how following the suggested actions could change practice versus what they would do themselves. In the interviews, this was actually the most common reply. The interviewees tended to want to evaluate the "resulting prediction" and its impact, rather than the rationale behind the prediction. Of course, we followed up on these questions by asking about those specifics (explainability), but interestingly this was rarely mentioned by the interviewees themselves. Still, we agree with reviewer A that this is much needed information, and we have revised the manuscript to at least include a quote showing that there is a need for explainability. Still, we would like to stress that this subject was rarely touched upon, and when it was discussed, it was only after the interviewers specifically asked about it. We added this quote:*"But if I would understand why it would predict a certain outcome, then I would be more inclined to consider whether I would or would not use it."* – I1. (line 389).

In addition, we added an additional result from the survey that had some connection to how the respondents feel their decisions can best be supported: *"When asked specifically how the participants would prefer AI algorithm outputs to be presented, there was no clear preference for absolute risk percentages (31/106; 29.3%), binary suggestions to take or not take a certain action (37/106; 34.9%), or risk categories (38/106; 35.9%)."*

Comment 4: 403 "Moreover, a trial phase in which physicians can test the AI algorithms and compare them to their own judgement, may further support implementation." I have big doubt whether event 100% accuracy of the AI algorithm in the trial would still be enough a clinician to trust it without knowing why it is making correct guesses. The paper should investigate this in more detail as it is critical.

Reply 4: we agree with reviewer A that it could be doubtable whether every clinician would fully trust an algorithm when a trial has shown 100% accuracy. However, based on our data, this is a conclusion we indeed can draw: in the results section, paragraph *Intervention characteristics – evidence strength* we explain that <u>all</u> interviewees stated that having a published trail would provide the best evidence, thereby enhancing their trust and facilitate implementation. However, of course a trial is not solely going to gain trust, more is needed, i.e. insight in how the algorithm is making the guesses (like reviewer A states). Therefore, in both the discussion and conclusion multiple strategies are discussed. In the discussion we stated: *This suggests that an international publication in itself is not sufficient for sustainable implementation. In addition, access to knowledge and information about the algorithm is essential.* We believe this is in line with reviewer A's comment that clinicians need to know more about the predictions of the algorithm.

Comment 5: Remark: Each trial costs. If there is no 'positive tension' (as defined in the paper) there will be (quite likely) no trial. Having no trial cannot show the relative advantage of using AI, so the problem could become a chicken and egg problem in this case. Thus, other ways that could facilitate the development of trust until reaching the point of the clinical trial and quantitative evaluation of the AI tool performance.
The big question that is missing in the subway is related to the comment I put for 295-196. Would a clinical trust (more) if he/she could see the factors driving the tool's decision, i.e. explanation of the prediction the algorithm had made? The authors should investigate this question.
Reply 5: We certainly agree with reviewer A that these concepts of tension and advantage are intertwined. However, we do not fully agree with the notion that this will inevitably result in a chicken and egg problem. In our opinion, clinical research will be the starting point of the AI/ML lifecycle. Researchers will identify problems that may not always be obvious to physicians (and thus there may not be tension yet) and try to create tools which to solve those problems. By showing these tools to physicians and convincing them of the added value (by showing relative advantage), a tension to change may be sparked. Local champions are indispensable in that regard. Of course, physicians need to feel that they can trust this tool, in order to fully achieve this relative advantage. Therefore, it is important to study and address barriers which prevent implementation at that point, as we studied here. Clinical trust and explainability seem to be important in that regard. However, as stated in the answer to comment 3, the interviewees in our study rarely addressed this item themselves. They would rather talk about how they would evaluate the clinical impact of such a tool compared to their own judgement, than talk about the importance of explainability. We did add a new quote on this aspect, which was one or the rare instances when this was mentioned (in response to a specific question by the interviewers): *"But if I would understand why it would predict a certain outcome, then I would be more inclined to consider whether I would or would not use it."* – I1. (line 389).

Reviewer B
The authors present a three-stage mixed methods study, building on their blood culture prediction model, to identify barriers and facilitators to AI implementation in healthcare. They use CFIR and ERIC to aid analysis and interpretation of their findings.

General Comments:
Comment 1: Your objective is to identify general insights about AI for clinical practice; however, the study is restricted to a particular geographic area, and the study participants do not represent all types of physicians or health care settings. The study has compelling results, but there is not strong justification for the study fundings to extend into other geographical regions, types of AI tools, and healthcare settings – the findings should be further contextualized to avoid

misrepresenting what is supported vs. what is presumed to generalize but requires further research to confirm.
Reply 1: The qualitative parts (= interviews and focus group) of our study were indeed 'restricted' to a single medical institution. However, this is not deemed restricting in the method used (qualitative research). The goal of qualitative research to study a complex phenomenon in-depth and from the perspective of the physicians themselves, as they are ultimately the end-users of algorithms. This is what we did in the qualitative parts of our study. Since we aimed to find broad and generalizable results, we also conducted the nationwide questionnaire (the quantitative part of our mixed-methods study). Although we fully agree with reviewer B that this can still be considered as a distinct geographic area, it already showed that the results were generally confirmed by this broader group of physicians. Furthermore, our results are quite similar to those by others in the literature (in different settings), which also suggests the results are generalizable to a wider setting. Still, we revised the wording of our discussion on this topic to avoid misrepresentation: *"Still, it would be helpful to tailor any implementation of an AI tool to the local context and end-users, for which additional surveys and interviews in those settings are needed to confirm the generalizability of our results."*

Comment 2: Result subsections tend to begin with a statement about "Physicians", followed by a quote etc. At some points you refer to participants or interviewed physicians – for clarity, any statements based on study participants should be specified as such, e.g., "interviewed physicians" to distinguish it from any information that is based on "general knowledge" or previous research (which should be referenced).
Reply 2: in the context of readability we tried to avoid too many duplication terms. However, we agree with reviewer B that this does not help readers in terms of interpretability. Therefore we changed this throughout the text of the results section; we now distinguish between 'interviewed physicians' and 'surveyed physicians'. (See tracked changes)

Comment 3: Some of the result subsections were included in the survey and others are not reported as such, e.g., inner setting – compatibility. Greater details about what was or was not included in the survey and why are needed.
Reply 3: Indeed, reviewer B is right that not all themes found in the qualitative part of the study were included in the survey. This is because we only included the most prominent constructs (identified during the qualitative interim analyses) in the survey. This was specified in line 225-226. Moreover, we made choices considering what themes were suitable for a quantitative survey, both in terms of topic and length of the survey. We aimed to keep the survey short and concise to ensure a good response rate. Readers can find the exact survey questions (and the included themes) in additional file 1. To further clarify this method, we added an additional sentence to the methods part about the survey: *"To keep the survey concise, we only incorporated questions on important topics identified in the interviews."*

Specific Comments:

Comment 4: Page 6, Line 107-109: is there a comparable statistic that could be used for Amsterdam to help readers better understand the relative state of AI implementation in the location where the study takes place?

Reply 4: We agree with reviewer B that it would be very helpful to share such statistics for Amsterdam or the Netherlands specifically. Unfortunately, these seem hard to find. We did find a reference which included information on AI-based tools approved by the European agencies. Since this information is more appropriate to understand the relative state of AI in the region where the study was conducted, we changed the reference. We now state: "*Up until 2020, only 222 AI tools were approved by the US Food & Drug Administration (FDA) and 240 in Europe (of which 124 in both)"*

Comment 5: Page 7, line 127: I appreciated the use of an existing, "real" project to guide discussions and examples, and the citation to the full piece of work is useful. Additional useful context would be to know whether at the time the interviews began your tool been implemented anywhere?

Reply 5: we agree that additional information would be useful. Based on this comment and other reviewers comments, we have moved the information from the introduction to the methods section. Moreover, we have added additional useful information regarding the work: *Then the topic narrowed down to a clinical case vignette about an AI blood culture tool to provide physicians with specific details, questions and prompts. This AI tool was recently developed by our research group, it predicts the outcomes of blood cultures in the emergency department, which may help avoid unnecessary testing and associated harmful effects. During the time of interview, focus group and survey (and to date) the blood culture tool was not implemented in clinical practice. We included this tool as clinical case vignette to provide interviewed physicians with real examples from a real project, and enhancing discussions.*

Comment 6: Page 10, line 187: Is something like "potential AI end-users" more accurate as using AI does not seem to be an eligibility criteria?

Reply 6: we agree with reviewer B, and have revised this accordingly.

Comment 7: Is there an estimate of the survey response rate?

Reply 7: For this type of nationwide survey, the local privacy regulations allowed us only to collect data through an anonymous link. Therefore, we were not able to track the amount of potential participant reached by the survey and we consequently cannot calculate a response rate. We do show the number of answers to all the individual questions in the results, indicating that most respondents answered all questions.

Comment 8: Table 2: There is only one respondent for several named Specialty groups - can the "Other" category be written out in full since it also only has one

respondent?
Reply 8: unfortunately, this is not information we have access to. The 'other' category was not followed by an open text field. Due to the anonymity of the survey we also cannot get this information in hindsight.

Comment 9: Page 14, Line 280-281: is "adaptable to patient population" referring to technical performance generalizability (e.g., accuracy) or related to workflow integration and the type of information that may or may not be useful to access in a clinical encounter (e.g., user interface)?
Reply 9: We thank reviewer B for this comment, as this as indeed unclear. We revised the text to include a brief explanation: *"An algorithm has to be adaptable to their patient population (regarding predictive performance), and be easy to integrate with existing workflows."*

Comment 10: Page 15, line 304: an example of a structural characteristic in this setting would be useful.
Reply 10: we agree with reviewer B that this is useful information, we added some examples in text: *"Structural characteristics (e.g. the social architecture or maturity of an organization)."*

Comment 11: Page 20, line 429: augmented clinical decision-making seems to be the focus of the study, but under "evidence strength" results the idea of cost-effectiveness/process optimization came up – would your results hold for these types of tools as well, or are some distinct considerations needed? This might be future work?
Reply 11: To address this question, we first would like to clarify that we believe some of these components to be following each other. We believe cost-effectiveness can be the result of augmented clinical decision-making, by making the most financially optimal treatment/diagnostic choices. The same can be true for process optimization, as the augmenting of clinical decisions may lead to a more streamlined diagnostic/treatment process with less redundancies. In our interviews, this was usually the way these relations were addressed, and that is how they were meant in this manuscript. The interviews and survey did not really assess specific tools for the optimization of, for example, laboratory processes. That could indeed be future work, although the authors consider most of these tools to be forms of automation, rather than artificial intelligence (which is of course a separate discussion).

Comment 12: Page 22, line 480: Consider whether other potential biases, like self-selection bias, may be present based on your recruitment methods.
Reply 12: we agree with reviewer B that this could have been a potential source of bias in the recruitment method of the survey. We have added this in the limitations section: *"Secondly, it could be possible that the survey was subject to self-selection bias, i.e. the physicians who chose to respond to the survey might have differed from*

the group of physicians that chose not to respond."

Reviewer C

I really enjoyed reading this article which adds urgently needed depth to the qualitative literature base and has a well-presented theoretically informed approach. I hope the revisions I suggest do not feel 'major', feel feasible within the review window and that the authors feel they add value.

Content comments

Comment 1: Line 111 - Replace the Benjamens et al reference as this simply contains the detail referenced in its introduction section not as a finding of the study. Much more contemporary data could be leveraged from https://aiforhealth.app/ and it's accompanying 2022 LDH write up by J Zhang et al.

Reply 1: We thank reviewer C for this excellent suggestion. At the moment of writing this manuscript, this paper was not yet available. It is a great reference to the point we are trying to convey, and we revised the manuscript to include it.

Comment 2: Line 107-109 – This is either contested or inaccurate which should be made clear with expanded references. Some authors (Lyell D et al 2021 BMJ Health and Care Informatics) have reported FDA approvals of AI enabled products as early as 2008 (IB neuro) and other authors (https://doi.org/10.1016/S2589-7500(20)30292-2) reviewed both European and American approvals which would broaden the relevance of your figures and also arrived at a much higher total of approvals.

Reply 2: These are both excellent suggestions by reviewer C. Since these papers make some contradicting claims, we opted to remove the statement about the earliest FDA approached tools altogether. We further changed the reference to include the second reference suggested by reviewer C, as it makes sense to include the paper with European approvals since these are closest to the place where the study was conducted.

Comment 3: Line 121 – Completely disagree that the CFIR is unique and I think it is counter-productive for the field to project the idea of a 'right'/best selection of theoretical approach. I think it's a good option and the justification lines 123-124 is great but could be expanded. E.g. why use a determinant framework, why not a process model? Why not use a technology focused determinant framework like NASSS or a less complicated one like TDF? It may also be worth using or referencing the 2022 update of CFIR published in Implementation Science.

Reply 3: we do believe that for the aims of our study CFIR (with ERIC) was the best fit and justified, which we will explain hereafter. However, we agree with reviewer C that 'unique' was a poor choice of wording, we changed this in the manuscript.

We appreciate reviewer C's effort to direct us towards other potential frameworks/models, and we looked into all of them to assess whether we could

use them:

- Process model: a process model aims to represent a certain process. Our study aimed to identify barriers and facilitators for implementation. Mapping/ visualizing a certain process was not our aim (especially since the algorithm has not yet been implemented), therefore we do not believe a process model to be a viable option for our study.
- NASSS: this seems like an interesting framework. However, the NASSS framework is only usable retrospective (i.e. after the implementation has already taken place). In our study, we aimed to identify facilitator and barriers pre-implementation. Therefore, we do not believe that NASSS could be a viable option for our study. Moreover, the NASSS is way less used in other studies (pubmed search provides 43 hits), and therefore less usable in terms of comparability.
- TDF: we agree with reviewer C that TDF would also have been a good choice of framework. Both TDF and CFIR are well-operationalized and theory-based implementation (determinant) frameworks. The TDF authors point out that their framework also direct to other relevant frameworks, like CFIR, for example for social influences and environmental context (which were factors in our study). We looked into the possibility of combining the two frameworks, but a systematic review by Birken et al (2017) has shown that combining the two is usually not justified and may reflect misleading wisdom. If one has to be chosen over the other, we deem CFIR to be a better fit for our aims, as it can be used in combination with the ERIC tool and thereby provides the opportunity to identify implementation strategies. Moreover, the TDF also seems to score a bit less in terms of comparability (568 hits in pubmed)

In conclusion, we stand by our choice for CFIR over other frameworks that reviewer C mentioned. Our study aimed to identify barriers and facilitators, CFIR provides a practical and theory/science based framework to do so. To date, CFIR had been cited 3.555 times in various implementation studies. Moreover, it can be used in combination with the ERIC tool, which provided more valuable insights.

Comment 4: Line 127-129 – If the blood culture example is a core part of the study, then the explanation of the intervention is far too thin. It is helpful that a technical validation article is represented, but even reading that it's unclear if the tool considered here is the logistic regression model (a technique that some would dispute as 'ML' and therefore 'AI' under certain definitions) which have been used in clinical practice for decades or the gradient boosted decision trees which holds more novelty if not significantly superior performance. I think it may be better to remove the reference to the blood culture work from the introduction all together, as it seems to me it was just used within the topic guide as a hypothetical example to stimulate responses and that as a pre-clinical tool no participants had practical experience of its use to draw on for their answers. Similarly, I'm not sure the survey data presented about whether or not the blood culture tool seemed like a good

idea is relevant. If the aim is to 'find general insights that could be applicable to 131 a wide variety of AI-tool implementations' then it seems odd to place such emphasis on a single potential use case which hasn't been directly experienced by participants.

Reply 4: The blood culture AI tool was not a core part of the study, but rather a clinical case vignette to guide discussion. We agree with reviewer C that the information should be deleted from the introduction, and better explained in the methods. Hence, we have added useful information regarding the work in the methods section, and clarified its purpose: "*Then the topic narrowed down to a clinical case vignette about an AI blood culture tool. to provide physicians with specific details, questions and prompts. This AI tool was recently developed by our research group, it predicts the outcomes of blood cultures in the emergency department, which may help avoid unnecessary testing and associated harmful effects. During the time of interview, focus group and survey (and to date) the blood culture tool was not implemented in clinical practice. We included this tool as clinical case vignette to provide interviewed physicians with real examples from a real project, and enhancing discussions.*"

Moreover, we agree that we should not refer to both publications: only to the gradient boosted decision trees (which was the tool that was used as the clinical case vignette).

Comment 5: Table 1 – add column describing participants experience with clinical AI tools as in table 2

Reply 5: unfortunately, this is not information we have access to, hence we cannot add it.

Comment 6: Methods 205 – This feels more like a framework analysis method to me so it would be good to have some references to support the 'deductive direct content analysis' to understand the approach better.

Reply 6: The framework analysis method that reviewer C mentions is a form of deductive direct content analysis. 'Direct content analysis' and 'deductive approach' are umbrella terms in qualitative research. For clarity regarding the approach we have added a reference to Green and Thorogood.

Comment 7: Results 221 – I find it slightly concerning that data saturation was reached after 10 interviews and a single focus group. If this was the case then I think it's important to think why that is in the discussion section – I'd suggest it may be due to a narrow sampling strategy of participants with little real-world experience to draw on and low diversity of perspective.

Reply 7: We respect reviewer C's concerns, However, there is no golden standard or sample size calculation for qualitative research, the n is based on data saturation. Saturation is reached when no new information is being generated from the interviews/focus group and analysis. When saturation is reached is dependent on your sample (homogeneous or heterogeneous), interview

(structured, semi-structured, or open) and aims (how tightly circumscribed is the subject). This means that n=10 could be sufficient, but in other studies n=20. In our study, saturation was reached with n=10 individual interviews and one focus group with n=5, because no new information and/or topics emerged. We can therefore conclude that this sample size was sufficient for our aims. Reaching the saturation after this n was not due to narrow sampling or any other reasons that reviewer C proposed; it was simple because no new topics emerged. Reaching the saturation 'soon' could have been due to the fact that we performed both individual interviews and a focus group, meaning we have the information on the topic that people are willing to share both 'publicly' and 'privately'.
Lastly, our study was mixed methods, meaning that we not only draw our conclusion based on the n=15 from our qualitative part, but also from our quantitative data that had a n=106. For these various reason we conclude that our sample size was sufficient, and not a subject of discussion in the limitations.

Comment 8: Results 225 – This reporting of 'most important potential benefit' seems to loose some of the potential detail that the survey design would offer. Is there a way to indicate when factors were also listed as 2nd most important? Perhaps in a bar chart?
Reply 8: As reviewer C suggests, there is indeed some information lost by presenting the data in this way. We felt that the most important item would be the most interesting, but we have now revised this to present the most prevalent ranks of these benefits:
"Surveyed physicians most often ranked these items in the following order of importance: patient outcomes, work process optimization, and cost-effectiveness (56/105; 53.3%). Some found costs to be more important than work processes, ranking them: patient outcomes, cost-effectiveness, and work process optimization (25/105; 23.8%). Among those who selected a ranking in which patient outcomes were not the most important potential benefit, the most common selection was: work process optimization, patient outcomes, and cost-effectiveness (17/105; 16.2%)."

Comment 9: Line 403 – I think this needs rewording, interviews are about identifying and exploring perspectives not quantifying their prevalence so I don't see the findings mismatched with those of the survey.
Reply 9: we agree with reviewer C that qualitative research (i.e. interviews and focus groups) are about exploring perspectives and not about quantifying. However, our study was not merely qualitative; it is mixed methods. The goal of mixed-methods is not solely to report quantitative and qualitative findings separately in one article; it is to integrate and combine findings. That is what we did throughout the results section. Therefore, we deem our original wording fitting for the (mixed-methods) approach that was used in this study.

Comment 10: Line 480 – The limitation section needs expansion. Many feel that one of the distinguishing features of AI implementation over other technologies is

its dependency on multi-stakeholder input so the exclusive recruitment of healthcare professionals is justifiable on feasibility grounds, but does not address the important gaps in the literature regarding other stakeholder perspectives. I would also draw attention to the (unavoidable) issue that a small minority of participants have any experience of AI in clinical practice and so their perspectives are important to understand, but are unlikely to have much basis and are likely to be transformed by real-world exposure when it occurs.

Reply 10: we agree with reviewer C that our initial limitation section needed expansion, we did so: "*Secondly, it could be possible that the survey was subject to self-selection bias, i.e. the physicians who chose to respond to the survey might have differed from the group of physicians that chose not to respond.*"

However, we believe the multi-stakeholder perspectives to fall beyond the scope of our study and aims. We aimed to identify facilitators and barriers to AI implementation for end-users. End-users will be physicians, and not wider stakeholders. Of course other stakeholders are important and could be relevant in other research, but it falls beyond the aims and scope of this study. Furthermore, we feel that the lack of experience among the participants is exactly what we needed in this study, as it reflects current clinical practice. We do agree that these perspectives will change by real-world exposure. However, the current study aimed to understand barriers that would withhold the participants from being open to such real-world exposure.

Minor comments

Comment 11: The word 'apprehended' is used in the abstract, discussion and the conclusion and feels awkward to me. Consider substitution or restructuring e.g. end-users cam see the potential value….. the value appears authentic to end-users…. Or similar

Reply 11: we agree with reviewer C and reworded to "acknowledge"

Comment 12: Line 80 of abstract 'Resulting, the ERIC tool displayed….' Odd wording, consider replacing resulting with consequently

Reply 12: we agree with reviewer C, and reworded accordingly

Comment 13: Line 164 – no > not

Reply 13: we reworded accordingly

Comment 14: Line 243 - lead > led

Reply 14: we reworded accordingly

Comment 15: Line 279 – 'To be facilitating…' reword

Reply 15: we reworded accordingly

Comment 16: Line 382 – none > no

Reply 16: we reworded accordingly

Comment 17: Fig 1 – mention in the legend that these codes are a-priori constructs from CFIR
Reply 17: we added this to the figure legend

Comment 18: Topic guide – typo 'Would you way of thinking' > your
Reply 18: we reworded accordingly

Reviewer D
The paper presents results from an interesting study on the barriers and facilitators of implementing Ai in clinical practice. The paper is well written. However, I have a few comments since some important details are missing.

Comment 1: Line 127ff: The authors mention that they developed an AI tool. It is unclear which role this tool plays. Later on it seems that it has been used somehow in the survey. But details are missing. Please make clear the role of the tool, how it was used for the study, which information the participants had on the tool, where they all users of that tool etc. If the tool is of importance for the study, a few more details on it would be helpful.
Reply 1: we agree with reviewer D that in the initial submission it was unclear what role the tool played, and that more details regarding the tool are needed. More of the reviewers had this advice. Hence, we move the information from introduction to the methods section, and have added useful information: "*Then the topic narrowed down to a clinical case vignette about an AI blood culture tool. to provide physicians with specific details, questions and prompts. This AI tool was recently developed by our research group, it predicts the outcomes of blood cultures in the emergency department, which may help avoid unnecessary testing and associated harmful effects. During the time of interview, focus group and survey (and to date) the blood culture tool was not implemented in clinical practice. We included this tool as clinical case vignette to provide interviewed physicians with real examples from a real project, and enhancing discussions.*"

Comment 2: Line 150 "The first few interviews...." - how many exactly?
Reply 2: we understand reviewer D's curiosity to exact numbers from a quantitative point of view. However, this is a mixed methods-study, and this data stems from the qualitative part of the study. In analyzing and reporting qualitative research it is not intended to quantify in exact numbers. Reporting numbers/ quantifying of qualitative output is deemed controversial. Therefore, (to adhere to qualitative research guidelines) we chose to stick to our original wording.

Comment 3: Line 186 "rank the most prominent barriers" - please provide more details. What was your rating scale? I assumed at that line, that the items were not described since they are results from phases 1 and 2. But later I recognized that authors never clearly state the items that are part of the survey (except in the appendix that shows the survey). Please provide more details on this.

Reply 3: we agree with reviewer D that some important details here are lacking. We reworded the section the be more clear, and added some details on the results. We added the following:
"Surveyed physicians most often ranked these items in the following order of importance: patient outcomes, work process optimization, and cost-effectiveness (56/105; 53.3%). Some found costs to be more important than work processes, ranking them: patient outcomes, cost-effectiveness, and work process optimization (25/105; 23.8%). Among those who selected a ranking in which patient outcomes were not the most important potential benefit, the most common selection was: work process optimization, patient outcomes, and cost-effectiveness (17/105; 16.2%)."

Comment 4: Line 226: "baseline characteristics" - I am not sure whether "Baseline" is the correct word here. With baseline, I would expect some ground truth where results are compared to. Maybe "demographic characteristics" is less misleading?
Reply 4: we agree with reviewer D that this was poor choice of wording, and reworded accordingly both in text and tables.

Comment 5: Result section: Authors are mixing up results from the three phases. I would suggest to separated phase 1 and 2 from phase 3. Then in becomes clear how you came to your survey questions. To support the reader in understanding, a list of all aspects at the beginning of the result section would be great.
Reply 5: We understand reviewer D's comment from a quantitative research point of view. However, our study was mixed methods, hence we analyzed and reported our results in a mixed methods style. The goal of mixed-methods is not solely to report quantitative and qualitative findings separately in one article; it is to integrate and combine findings. That is what we did throughout the results section: with the qualitative (phase 1 and 2) and quantitative (phase 3) parts integrated. Therefore, (to adhere to reporting guidelines and the aims of our mixed-methods study) we chose to keep our original results section.

Comment 6: The number of n survey participants seems to change several times (line 255 and line 405 n=105, line 341 n=104, but at other places n=106) This has to be corrected (of is it is correct, authors have to explain this)
Reply 6: reviewer D notices right that the n is variable between questions within the survey, this is due to a difference in missing per question. We have further clarified this in the text and table: *"Notably, some questions were not answered by all participants. The number of answers and total number of responses are presented with all results."*

Comment 7: Line 341: Obviously, the survey participants were asked about the tool. But did they used it before? or which information did they got?
Reply 7: we have added this information: "*During the time of interview, focus group and survey (and to date) the blood culture tool was not implemented in clinical practice."* Moreover, appendix 2 shows the exact survey (and thereby the

information the survey participants had).

Comment 8: In the discussion of the limitations, I am missing sentences on whether n=106 is a representative amount. How many physicians would have been achievable with the procedure authors applied?
Reply 8: In the previous version, we did discuss the sample size of 106 shortly in the methods section. We agree with reviewer D that this is information that also needs to be addressed in the limitations section, where we now added a sentence to explain that we aimed to include 100 participants to ensure we had some variety. Since we did not plan any statistical tests, we did not do a power calculation. The revision in text: *Lastly, in our survey we included n=106 participants. We did not perform an a-priori sample size calculation, as we did aim to perform statistical testing. It is therefore challenging to make a statement regarding the representativeness of our sample size. However, we do believe that this sample size is sufficient to ensure a range of variety in the participants.*