PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Study protocol for a prospective, double-blinded, observational study investigating the diagnostic accuracy of an app-based diagnostic health care application in an emergency room setting: the eRadaR-trial
AUTHORS	Faqar-Uz-Zaman, S. Fatima; Filmann, Natalie; Mahkovic, Dora; von Wagner, Michael; Detemble, Charlotte; Kippke, Ulf; Marschall, Ursula; Anantharajah, Luxia; Baumartz, Philipp; Sobotta, Paula; Bechstein, Wolf; Schnitzbauer, Andreas

VERSION 1 – REVIEW

REVIEWER	Kiran Grant Canada
REVIEW RETURNED	21-Sep-2020

GENERAL COMMENTS	It was somewhat unclear how the hospital visits will relate to patient population as described. The authors should mention up front that visit 2 and 3 would only occur if needed, and that patients who are discharged from the ED can still be included in the study. The authors should also describe how follow up would occur in a patient immediately discharged from the ED, and how this relates to their attempt to assess the AI technology. I think this is important because a large proportion of all patients presenting with abdominal likely will not be admitted. Minor grammatical errors throughout

REVIEWER	Marta Fernandes MGH Boston, Massachusetts, USA
REVIEW RETURNED	26-Sep-2020

GENERAL COMMENTS	Page 23 - Correct 'Docotor' in the diagram

REVIEWER	Tadahiro Goto The University of Tokyo, Japan
REVIEW RETURNED	01-Oct-2020

GENERAL COMMENTS	General comments ========== I appreciate the opportunity to review this protocol. This protocol is well-written, and the study design is appropriate. 1. Please clarify what kind of information is collected through the Ada-APP? Only information on patient demographics and interviewed-data? 2. What will you treat patients who visit ED with abdominal pain and

	other chief complaints? (abdominal pain as a part of multiple complaints). 3. Please indicate a figure for RCT patient flow.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:
Comment to the Author:
"It was somewhat unclear how the hospital visits will relate to patient population as described. The authors should mention up front that visit 2 and 3 would only occur if needed, and that patients who are discharged from the ED can still be included in the study. The authors should also describe how follow up would occur in a patient immediately discharged from the ED, and how this relates to their attempt to assess the AI technology. I think this is important because a large proportion of all patients presenting with abdominal likely will not be admitted."
Minor grammatical errors throughout."
Answer to the reviewer:
We thank the reviewer for carefully evaluating our manuscript. In our revision, we have described the schedule of the visits more precisely.
In Table 1, we have mentioned in the footnote, that visit 2 and 3 are left out, if the patient is discharged before. For patients, who will be immediately discharged from the ED, only visit 1 and visit 4 will be performed, obtaining data about baseline assessment, App-interview, symptoms, diagnostics and therapies. If patients are admitted to the hospital after presenting to the ED, visit 2 and visit 3 will be performed additionally, depending on the length of hospital stay (visit 2 on day 7 and visit 3 on day 14). During these visits, data about further diagnostics, complications and therapies will be collected.
In conclusion, we plan to include patients, who will be immediately discharged from the ED on the same day, who will definitely form the larger proportion of patients, as well as patients who will be admitted to the hospital after presenting to the ED. This information has been added to the revised manuscript (Page 6 and 8).
However, all patients included in our study will be followed up on day 90 (visit 5), performing a structured telephone interview, irrespective whether they have immediately been discharged or have been admitted to the hospital.
"Minor grammatical errors throughout."
The paper underwent English editing and the whole text has been rephrased. Correction in misspelling and rephrasing have not been marked.

Reviewer 2:
Comment to the Author:
"Page 23 - Correct 'Docotor' in the diagram"
Answer to the reviewer:
We apologize for this mistake. 'Doctor' has been corrected.

Reviewer 3:
Comment to the Author:
"1. Please clarify what kind of information is collected through the Ada-APP? Only information on patient demographics and interviewed-data?
2. What will you treat patients who visit ED with abdominal pain and other chief complaints? (abdominal pain as a part of multiple complaints).
3. Please indicate a figure for RCT patient flow. "
Answer to the reviewer:
1. We thank the reviewer for this comment. The Ada-App® obtains only information about patient demographics, patient history and information about current symptoms using a structured algorithmic

pathway of questions based on a medical knowledge base and AI-based machine learning system. No further information are collected from the patient. This has been clarified in the revised version (Page 6).
2. We thank the reviewer for this interesting question. We have thought about this problem as well. This addresses a problem, we indeed have to face in the ED in reality. However, we will only include patients, who visit the ED with abdominal pain as a chief complaint, even though, the patient might have other chief complaints as well. Since our study aims to evaluate the performance of an App-based decision support system in real clinical settings, these cases should be evaluated as well in our opinion. Nevertheless, it might well be, that the final diagnosis is not related to an abdominal disease.
3. As this is an observational study, no randomization will be performed. A study flow chart of our study is displayed in Figure 1.

VERSION 2 – REVIEW

REVIEWER	Kiran Grant University of Toronto, Canada
REVIEW RETURNED	26-Nov-2020

GENERAL COMMENTS	Clearer description of the underlying functioning of the Ada applications as well how it may compare to other similar applications is needed. Further, the level of detail in the assessment is unclear. For example, dizziness may be commonly reported symptom, but for this to be clinical useful it must be determined whether this description is meant to refer to lightheadedness/pre-syncopal symptoms versus vertigo. If the app is unable to make such distinctions I think it would be difficult for it to make any accurate assessments, and this would also undermine the quality of the input data as the underlying AI would be working with incomplete or inaccurate data (with the app assuming all dizziness is lightheadedness or vertigo or neither). As we know with any such algorithm, it is only as good as the data inputted into it. Second, It is not clear that earlier diagnosis is the key to improving ED throughput. Instead, wouldn't targeting the application to distinguish key ED zoning, advanced imaging needs, and disposition decisions be most useful? For any such patient with abdo pain, whether the algorithm says its an appendicitis or not based on the patient describing the classic history (or some variation therein) would not be as useful in the ED setting as the same algorithm being able to accurately predict whether this patient can be seen in the ambulatory or acute side of the department, if they will need advanced imaging such as CT, and whether they are likely to be admitted or sent home (and if they are admitted, to what consultant service). I appreciate the the authors may be suggesting that the diagnosis would indirectly answer these same questions, however I think it is a much taller order to train an Ai to accurately distinguish between the multitude of potential diagnosis of abdo pain on history alone than it would be to focus it on key, often binary decisions that ultimately determine the patient's flow through the department. Predicting such flow up front could enable greater coordination between staff and ED resources, and improved predictions of ED wait times and instances of access block

VERSION 2 – AUTHOR RESPONSE

thank you very much for the feedback and the additional comments from Kiran. Obviously, we have tackled an interesting topic here. We do understand the new questions and do hope that we can be clarify them accordingly.

1. Clearer description of the underlying functioning of the Ada applications as well how it may compare to other similar applications
is needed.

Answer to reviewer: thank you very much for the comment. The rationale why we used the Ada app was indeed not well elaborated. A conclusive and clear statement has been added in the patients and methods section of the manuscript. “The Ada app is a class I medicinal product certified in accordance with theDIN ISO 13485. In this study we investigate this solution and do not compare it to any other solution of the market. This is completely out of our scope. We have decided to explore this software solution in a doubleblinded
fashion due to the fact that publications on the software outpower other solutions in a convincing evidence-based way as several trials of the company show. The data for that are displayed on their homepage (JMIR Hum Factors (2020) doi: 10.2196/19713., Ann Rheum Dis (2020) doi: 10.1136/annrheumdis-2020-217125., Orphanet J Rare Dis (2019) doi: 10.1186/s13023-019-1040-6., medRxiv (2020) doi: 10.1101/2020.05.07.20093872., medRxiv (2020) doi: 10.1101/2020.06.16.20126466., medRxiv (2020) doi: 10.1101/2020.06.19.20135590., Zeitschrift für Rheumatologie (2020) doi: 10.1007/s00393-020-00834-y., medRxiv (2020) doi: 10.1101/2020.07.07.20147975..
Ada is a free-downloadable certified medicinal device and has been validated in different studies by the marketing authorization holder and developer team. It has shown a higher accuracy (73%) in comparison to other apps (38%) when compared to the correctness of symptom checking. The App was superior to other apps when the hitlist of the 5 most probable diagnosis were compared (84% vs. 51%).

The algorithm is not an official source document and the source codes are not a public document as this is the intellectual property of the company. However, the evidence shows that the algorithm is superior to other solution on the market, it has been validated by the
company, and the data were the basis for the certification as a medicinal product class I (CEmark in accordance with DIN ISO 13485).”

2. Further, the level of detail in the assessment is unclear. For example, dizziness may be commonly reported symptom, but for this to be clinical useful it must be determined whether this description is meant to refer to lightheadedness/presyncopal symptoms versus vertigo. If the app is unable to make such distinctions I think it would be difficult for it to make any accurate assessments, and this would also undermine the quality of the input data as the underlying AI would be working with incomplete or inaccurate data (with the app assuming all dizziness is lightheadedness or vertigo or neither). As we know with any such algorithm, it is only as good as the
data inputted into it.

Answer: please see answer 1. I guess that an authority cleared medicinal device with the particular CE-mark is a sufficiently tested solution. Otherwise, it would not be allowed to be used on the market for the intended purpose. Please, do understand that we are not the developers of this software. We are clinical users, and we made the decision to use this software solution based on their convincing data. Those data are also the regulatory basis for clearance by authorities.

3. It is not clear that earlier diagnosis is the key to improving ED throughput Answer: exactly, this is

our hypothesis that we investigate. We try to find out, if an earlier diagnosis leads to a better treatment of patients or if this is not necessary. This is all objective of our clinical trial and cannot be answered at this point.

4. Instead, wouldn't targeting the application to distinguish key ED zoning, advanced imaging needs, and disposition decisions be most useful? For any such patient with abdo pain, whether the algorithm says its an appendicitis or not based on the patient describing the classic history (or some variation therein) would not be as useful in the ED setting as the same algorithm being able to accurately predict whether this patient can be seen in the ambulatory or acute side of the department, if they will need advanced imaging such as CT, and whether they are likely to be admitted or sent home (and if they are admitted, to what consultant service).

Answer: This may be true in other environments and structures. For sure it is not true in most German settings. Here, a deeper knowledge of the tracking of patients in our ED is required. Every patient that has abdominal pain is seen in the ED. There is no referral to an ambulatory or the acute side of a department. The patients stay in this environment until a decision is made how to proceed with a diagnosis and a treatment plan. The acute sides of any medical department in most German hospitals is the emergency room! This is where we have a large interdisciplinary setting where patients are treated as required. Abdominal pain in this context can be tracked either to internal medicine, abdominal surgery, urology, gynecology, vascular surgery, or even cardiology etc. with close cooperation to the radiology department. The patient however will not be referred to any of those departments until a decision is made and has been seen as required.
However, your suggestions are already a potential answer and discussion point to our findings in a future manuscript of the data that we try to obtain. I do really appreciate these comments, but it is too early to talk about this.

5. I appreciate the authors may be suggesting that the diagnosis would indirectly answer these same questions, however I think it is a much taller order to train an Ai to accurately distinguish between the multitude of potential diagnosis of abdo pain on history alone than it would be to focus it on key, often binary decisions that ultimately determine the patient's flow through the department. Predicting such flow up front could enable greater coordination between staff and ED resources, and improved predictions of ED wait times and instances of access block.

Answer: thank you very much for this comment. You are once more discussing our potential results already. Obviously, you like the topic. This is our research question in a more elaborated way of answering. Right now, it remains speculative and we need to collect the data first. Basically, we do not train the AI. This is impossible for us. We use a certified software of an adequately trained AI algorithm and it has been trained and cleared by authorities, so we have to rely on the appropriate training the algorithm has in its intended use. I hope you do understand this. This is what we are trying to find out. Concerning the patient flow, the University Hospital Frankfurt/Main was one of the first hospitals in Germany that was certified in accordance with the DIN ISO 9001 for dedicated quality and risk management of the complete institution on all levels. Furthermore, we are a completely digitalized institution that tries to improve step by step on an already high level. The research question to explore potential improvements by using certified software aims at improving a system that is already very much standardized. From a management perspective there is no need to challenge or improve the basic workflow currently as proposed by you.