Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220386

User Satisfaction with an AI System for Chest X-Ray Analysis Implemented in a Hospital's Emergency Setting

Diego RABINOVICH[a], Candelaria MOSQUERA[a,b,1], Pierina TORRENS[a], Martina AINESEDER[c] and Sonia BENITEZ[a]
[a] *Health Informatics Department, Hospital Italiano de Buenos Aires, Argentina*
[b] *School of Engineering, Universidad Tecnológica Nacional, Argentina*
[c] *Radiology Department, Hospital Italiano de Buenos Aires, Argentina*

Abstract. The acceptance of artificial intelligence (AI) systems by health professionals is crucial to obtain a positive impact on the diagnosis pathway. We evaluated user satisfaction with an AI system for the automated detection of findings in chest x-rays, after five months of use at the Emergency Department. We collected quantitative and qualitative data to analyze the main aspects of user satisfaction, following the Technology Acceptance Model. We selected the intended users of the system as study participants: radiology residents and emergency physicians. We found that both groups of users shared a high satisfaction with the system's ease of use, while their perception of output quality (i.e., diagnostic performance) differed notably. The perceived usefulness of the application yielded positive evaluations, focusing on its utility to confirm that no findings were omitted, and also presenting distinct patterns across the two groups of users. Our results highlight the importance of clearly differentiating the intended users of AI applications in clinical workflows, to enable the design of specific modifications that better suit their particular needs. This study confirmed that measuring user acceptance and recognizing the perception that professionals have of the AI system after daily use can provide important insights for future implementations.

Keywords. user satisfaction; artificial intelligence; clinical decision support systems; radiography; chest.

1. Introduction

The acceptance of AI-based systems by health professionals is crucial to obtain a positive impact on the diagnosis pathway [1]. Understanding why specialists accept or reject a technology is necessary to better predict, explain, and increase user acceptance [2]. The evaluation of user satisfaction can help to identify the system's strengths and weaknesses and guide the planning and design of adequate improvements [3,4].

The primary goal of this study was to obtain a preliminary analysis of user satisfaction with an AI-based system for the automated detection of findings in chest x-

[1] Health Informatics Department, Hospital Italiano de Buenos Aires. 4190 Perón St., C1199AAB, Ciudad Autónoma de Buenos Aires, Argentina. Tel (+54) 01149590200 - ext. 5056; E-mail: candelaria.mosquera@hospitalitaliano.org.ar.

rays, named TRx, which was developed and validated at a health center. In this study, we evaluated the TRx application integrated in the Electronic Health Records (EHR) and the Radiology Information System (RIS) of our center. Our objective was to find patterns in perceptions that were common across users, and identify which factors are implied in the positive uptake of an AI-system for medical imaging, stratifying the results by users' specialties.

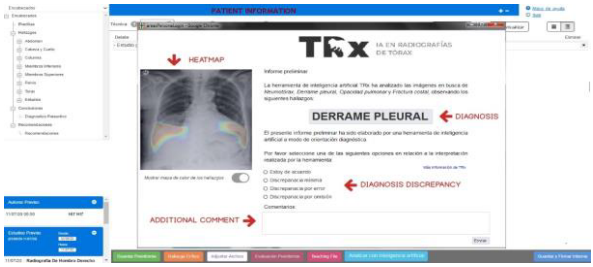


Figure 1. TRx interface for RIS users. Red captions indicate the main sections.

2. Methods

This was an IRB-approved observational mixed study (N° 6025), conducted at the Emergency Department of a 650-bed university hospital in Buenos Aires between January 1st and May 31th 2021. A total of 7689 chest x-ray studies were processed by TRx during the study period (average of 51 studies per day). Participants were selected as TRx intended users in the emergency setting: radiology residents and emergency physicians. To assess user satisfaction we evaluated the four factors of the Technology Acceptance Model [7]:

- **Actual system use**: the degree to which a person uses the technology.
- **Perceived usefulness**: the degree to which a person believes that using a particular system would enhance his or her job performance.
- **Perceived ease of use:** the degree to which a person believes that using a particular system would be free from effort.
- **Output quality**: the perceived correctness of the application´s prediction.

TRx is an AI application that assists users in chest x-ray interpretation. It combines four deep learning models that were trained for the detection of four critical findings: pneumothorax, rib fracture, pleural effusion and lung opacities [5]. In the interface, user feedback can be optionally completed at the time of image evaluation. TRx has two different questionnaires to address two distinct intended uses.

(1) When accessed through the RIS (radiology specialists), the questionnaire is focused on diagnostic performance: it consists of a four-point scale on the level of discrepancy with TRx's diagnosis, which was already used for evaluation of the radiology residents' preliminary reports [6] (Fig. 1).

(2) When accessed through the EHR (emergency physician or other specialties), this consists of a five-point Likert scale on the level of usefulness for their work.

3. Results

Regarding actual system use, we retrieved quantitative data on system use from TRx's database, which records the number of times someone accessed the application and found that the interface was accessed in 15.4% of studies (n=1186), with an average of 8 accesses per day. To estimate output quality, we considered the radiologists' feedback: in RIS questionnaires, the proportion of agreement varies greatly among images where TRx detected findings and those where it detected no findings: 90% and 34% respectively (Fig. 2b). Regarding perceived usefulness, we observed that 60% of answers classified TRx with a good utility level in EHR questionnaires (Fig. 2a).

Perceived ease of use was measured through a validated survey using the System Usability Scale (SUS), a widely adopted method for assessing user experience [8,9] (Fig. 3). The survey was answered by 13 professionals: 62% from the Radiology Department and 38% from the Emergency Department. It showed that most users agree that TRx is comfortable and easy to understand, requiring no technical support and no special training to start using it. The greater variation in questions about confidence and consistency suggests these might have been understood as referring to diagnostic performance.

Qualitative analysis was performed by interviewing participants who volunteered, following a structured list of questions. Six physicians were interviewed: three emergency physicians and three radiology residents.

Actual system use: In general, all interviewed participants agreed they would continue using this tool and they would recommend it. An emergency physician noted that she often evaluates X-ray images directly in the portable equipment at the patient's bedside instead of using the EHR, so TRx is not used in those cases. Radiology residents used TRx through the RIS during their X-ray training period, analyzing about 20-30 chest x-rays per day.

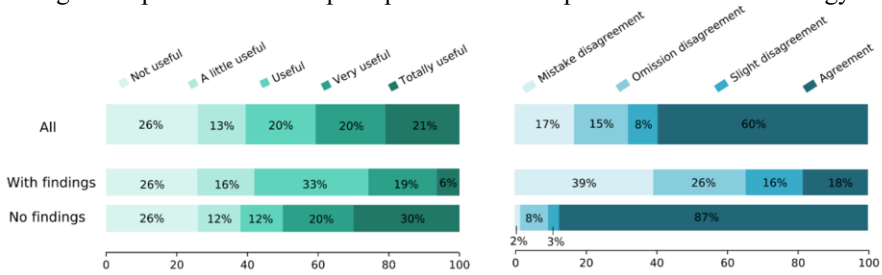
Perceived usefulness: all participants agreed they find the system useful for their daily work, and stated they look at the original image first, to make their own diagnosis, and only after that they look at TRx output as a second opinion. We identified two common TRx uses mentioned as helpful by all interviewed emergency physicians: to confirm a finding has not been omitted, and as an alert of a critical finding, particularly when the patient shows no suggestive clinical signs. Radiology residents also agreed on the utility as diagnosis confirmation, particularly in images with no findings. Additionally, they mentioned TRx's utility to reduce human errors associated with fatigue or rush caused by work overload.

Perceived ease of use: all participants rated the system usability as very good, describing the interface as easy to understand and practical. Suggested improvements included adding tags to identify specific findings in the heatmap and providing visualization tools on the heatmap, such as zoom and scroll.

Output quality: emergency physicians perceived a high output quality, but admitted it might be due to limited use. When comparing it to their own diagnosis, they found they agreed with TRx almost always. However, radiology residents found that diagnostic performance varied greatly across radiological findings. They all concur that the system is especially good at detecting pneumothorax and pleural effusion, and also experienced a good agreement with TRx in normal images. However, they perceived a poor performance for lung opacities, with many false positives. In addition, they noticed a decreased performance in chest x-rays with poor acquisition technique, for example studies from patients who were lying in bed.

4. Discussion

In this work we present the preliminary results of a study on user satisfaction with an AI application for chest x-ray diagnosis implemented in real clinical practice. We found general patterns in users' perceptions on four aspects from the Technology



(a) (b)

Figure 2. User feedback collected from 401 answers. (a) EHR answers. (b) RIS answers

Acceptance Model. The best perceived aspect was the ease of use, which was consistently remarked as very good both in the survey and interviews. Perceived usefulness focused on using TRx to confirm no omission of findings, while emergency physicians also mentioned its utility as an early alert on potential critical findings and radiology residents expressed that TRx guidance could help reduce human diagnostic errors.

A difference among specialties was also observed in the perception of output quality, as physicians generally expressed that TRx has good detection accuracy, while radiology residents provided a detailed judgement over different pathologies, which is aligned with the quantitative results. This confirms that the expected utility of TRx is different in the Emergency and Radiology Departments. This should be considered when adjusting TRx implementation, adapting the system to suit the intended use in the EHR and the RIS respectively.

This study confirmed that output quality is a decisive factor in user satisfaction. A system with good diagnostic performance is the first step to build users' trust in its output, which is certainly required to increase use during daily practice. In particular, results showed that an essential step in TRx future versions should be increasing the performance for lung opacities. We also confirmed that system use is improved by a suitable placement of the application throughout clinical workflows, which was better achieved for the radiology workflow than the emergency one.

The main limitation of this report is its small sample size, which impedes quantitative comparisons with precise statistical results. However, our data shows clear trends and interesting patterns that may guide further efforts in the development and clinical implementation of AI-based diagnostic tools. Future work includes collecting new participants to increase the significance of results.

The main strength of this study is that it reports on data from a real clinical implementation of AI in medical imaging diagnosis. User satisfaction with health AI applications has not been studied widely, and few prior works report on this topic [10-12]. Future research should focus on identifying bottlenecks and barriers that are met by different groups of users when using the system, to allow relevant modifications that could actually improve the system’s clinical utility in real healthcare settings [13].

5. Conclusion

This study of user satisfaction with TRx application yielded positive evaluations from participants. Emergency physicians and radiology residents shared common patterns regarding their perception of TRx's usability, while differing on output quality and usefulness for their work. These results highlight the importance of clearly

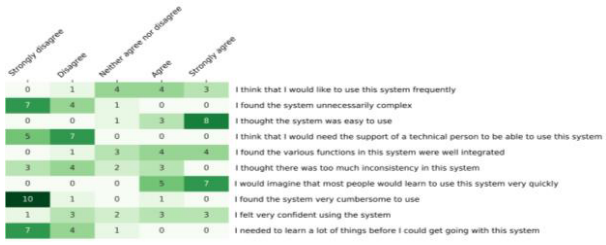


Figure 3. The ten questions on System Usability Scale and the count of survey answers.

differentiating the intended users of AI applications in clinical workflows, to allow a separate analysis of their satisfaction and to adapt specific designs that meet their needs. Measuring user acceptance and recognizing the perception that professionals have of the AI system after daily use can provide important insights for future implementations.

References

[1] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nature medicine. 2019;25(1):30-36.
[2] Høstgaard AM, Bertelsen P, Nø0r C. Methods to identify, study and understand end-user participation in HIT development. BMC medical informatics and decision making. 2011;11(1):1-11.
[3] Akhloufi H, et al. A usability study to improve a clinical decision support system for the prescription of antibiotic drugs. PloS one. 2019;14(9):e0223073.
[4] Kastner M, Lottridge D, Marquez C, Newton D, Straus SE. Usability evaluation of a clinical decision support tool for osteoporosis disease management. Implementation Science. 2010;5(1):1-12.
[5] Mosquera C, et al. Chest x-ray automated triage: A semiologic approach designed for clinical implementation, exploiting different types of labels through a combination of four Deep Learning architectures. Computer Methods and Programs in Biomedicine. 2021;206:106130.
[6] Diaz FN, Ulla M. Validation of an informatics tool to assess resident's progress in developing reporting skills. Insights into imaging, 2019;10(1):1-10.
[7] Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. Journal of biomedical informatics, 2010;43(1):159-172.
[8] Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction, 2008;24(6):574-594.
[9] Laugwitz B, et al. Construction and evaluation of a user experience questionnaire. In Symposium of the Austrian HCI and usability engineering group. Springer, Berlin, Heidelberg. pp. 63-76.
[10] Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean journal of radiology. 2019;20(3):405.
[11] Beede E, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI conference on human factors in computing systems. p. 1-12.
[12] Jauk S, Kramer D, Avian A, Berghold A, Leodolter W, Schulz S. Technology Acceptance of a Machine Learning Algorithm Predicting Delirium in a Clinical Setting: a Mixed-Methods Study. Journal of medical systems. 2021;45(4):1-8.
[13] Magrabi F, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications: a position paper from the IMIA technology assessment & quality development in health Informatics working group and the EFMI working group for assessment of health information systems. Yearbook of medical informatics. 2019;28(1);128.