Human-AI interactions in a trial of AI breast cancer diagnostics in a real-world clinical setting

Human-AI interactions in a trial of breast cancer diagnostics

Emma Engström, PhD

Institute for Futures studies, Stockholm, Sweden, emma.engstrom@iffs.se

KTH, Royal Institute of Technology, Sweden

Fredrik Strand, MD, PhD

Karolinska Institutet, Solna, Sweden, fredrik.strand@ki.se

Pontus Strimling, PhD

Institute for Futures studies, Stockholm, Sweden, pontus.strimling@iffs.se

This study discusses the design of a trial of AI breast cancer diagnostics in a real-world clinical setting at Capio St Göran Hospital in Stockholm, Sweden. We describe our approach to analyzing human-AI interactions in this context, relating the characteristics of the trial to recent research within AI implementation and AI ethics. Important concerns include ensuring patient safety and avoiding clinician overreliance on AI assessments. The AI will function as a third independent radiologist in the first screening stage, which is expected to increase sensitivity of the process while safeguarding the safety of the individual and maintaining trust in her relationship to the health care practitioner. We plan to evaluate the radiologists' confidence in the system over time and in relation to a potentially odd assessment by the AI. This discussion is expected to illustrate complications and possibilities related to 'the last mile' of AI in health care.

CCS CONCEPTS • Applied computing ~ Life and medical sciences ~ Health care information systems • Human-centered computing ~ Human computer interaction (HCI) ~ Empirical studies in HCI • Human-centered computing ~ Human computer interaction (HCI) ~ HCI design and evaluation methods ~ Field studies

Additional Keywords and Phrases: AI-Health, Breast cancer diagnostics, AI ethics, AI implementation

1 INTRODUCTION

1.1 The 'last mile' of AI in health care

While adoption of AI has been supersonic in some environments, such as mobile apps (Engström & Strimling 2020), its implementation in health care remains slow. This is particularly unfortunate within medical imaging, a pillar of medical treatment, because many aspects of radiology could be improved by AI-methods (Hosny et al. 2018). The expectation is that AI would enable more consistent and potentially more cost-effective decisions than human health care practitioners. However, several open questions remain in the field, for instance related to the nature of the interface between the AI and the radiologists, and ambiguity in terms of who is responsible in the presence of AI (Hosny et al. 2018). Sociotechnical processes may hence contribute to explain the slow implementation of AI in health care. It is a challenge to introduce autonomous systems into practices where tolerance for errors is small or non-existent, and trials are inherently exposed to errors. Clearly, it is immensely important to analyze potential mistakes within adoption of AI; otherwise, public trust may be undermined, and this may increase opportunity costs in the future, i.e., costs for missing out of potential benefits when choosing something else (Morley et al. 2020). Floridi (2019) argued against being overly cautious to AI in health, while stressing the need to be mindful of its potential ethical impacts. This study examines such impacts in relation to a real clinical setting.

1.2 Socio-technical processes within AI in health care

A challenge related to the 'last mile' of AI implementation in health care is that the AI does not act on its own, but that it needs to have a substantial impact on real-world processes to be meaningful (Coiera 2019). Coiera (2019) argued that AI development should be seen as an agile and iterative process, and that it is important to find a balance between using general AI technology and identifying local solutions (Johannessen & Ellingsen 2009). For illustration, IBM's Watson system for preventive mammography screening that was employed by the Capital Region of Denmark in 2017 was found to need further development before clinical use (Spielkamp 2019). One reason for this was that the system agreed with doctors in only 27% of treatment suggestions, which was attributed to the fact that it had been trained to follow American practices (Spielkamp 2019). This demonstrates unexpected problems that may arise as AI is implemented in a specific clinical environment.

1.3 Breast cancer screening in Sweden

In Sweden, population-based breast cancer screening was established in the 1980s. The national guidelines stipulate regular screening of women from age 40 to 74 years by inviting them to mammography at least every second year. Around 75% of women take part in this free program, which has been shown to decrease breast cancer mortality by up to 41% (Duffy et al. 2020). Already in the 1990s, software was developed to help radiologists in screening, however it suffered from a high rate of false positives, more than 1 per image, leading to radiologists spending much more time on each exam without finding a significant amount of additional cancer.

Not long ago, the development of AI in the form of deep neural networks has showed encouraging results, and many new commercial players are entering the arena. A recent study reported that a system developed by IBM based on deep learning could assess breast cancer at the same level as radiologists and may reduce the number of missed diagnoses (Akselrod-Ballin et al. 2019). Comparing three commercially available AI computer-aided detection algorithms in prospective clinical studies, Salim et al. (2020) reported that one of them performed at the same level as a radiologist. The combination of a first reader radiologist with the best algorithm resulted in higher positive cases for cancer than the combination of first and second readers, suggesting that it had sufficient performance to be evaluated as an independent reader in future trials (Salim et al. 2020).

This study describes our involvement in one of the first large-scale implementations of AI in screening mammography, 'ScreenTrust CAD'. The research is conducted at the Breast Imaging Unit at the Capio St Göran Hospital in Stockholm, Sweden. It is currently in the initial phase. The evaluation will include 55,000 women during a period of up to two years. The principal investigator is Fredrik Strand, breast radiologist and researcher at Karolinska University Hospital. Here we present our approach to studying human-AI interactions in this trial.

2 CASE STUDY OF AN AI-HEALTH TRIAL AT CAPIO ST GÖRAN HOSPITAL

2.1 The clinical trial

In the conventional screening procedure, radiology nurses acquire two mammography images of each breast, which are then screened by two radiologists independently. If any of them finds something potentially malignant in the images, the exam proceeds to a second reading, denoted a consensus discussion. In this dialogue, two radiologists evaluate the mammograms together to decide whether to recall the woman for further diagnostics. Of 1,000 screened women, around 80 cases are flagged for a subsequent consensus discussion, about 25 are recalled, and around five are eventually diagnosed with breast cancer. Because of the high number of mammograms to analyze, radiologists are in short supply in Sweden, and the vision is that the AI will eventually replace one of the two radiologists involved in the initial screening.

However, in the trial the AI is implemented as a third independent reader in the first screening in order to ensure patient safety whilst allowing for subsequent analyses of different screening scenarios. The AI assesses the images to generate a risk score and a binary decision for each exam: whether the woman should be considered healthy or go to the consensus discussion. A consensus discussion is initiated if an exam has been flagged by the AI or one of the radiologists. The AI's involvement in the consensus discussion is limited to the radiologists reviewing the AI results, including the malignancy score and a heat map describing where in the image it found the most suspicious area. Then, it is up to the two radiologists to make the final decision.

2.2 AI-implementation challenges

Coiera (2019) identified three challenges related to AI's adoption in health care, discussed below.

2.2.1 Measurement

Evaluations of an AI's performance need to shift from measuring its technical accuracy to assessing its impacts on people and processes. High technical accuracy in a research setting is a requirement, but it does not guarantee impact in practice. In the trial, we plan to address this by not only measuring the performance of the AI-system in terms of sensitivity and specificity, but also by assessing whether the performance of the entire screening procedure is enhanced as the technology is introduced. Further, we will examine changes in the interactions between the radiologists and the AI-system to assess whether adoption and/or trust may follow a K-shaped pattern over time, in which radiologists tend to increasingly rely on the system or increasingly disregard it; this way, we will learn whether the AI induces changes in attitudes and practices.

2.2.2 Generalization and calibration

This challenge relates to the fact that AI has been trained in a particular historical context and in one type of population, which may not match the context and the population where it is applied (Coiera 2019). For instance, there may be differences in the frequency and type of events between different populations (Coiera 2019). Notably, breast cancer rates vary across different demographics, with lower incidence and higher mortality among Hispanic, Asian, African American, and Native American women as compared to non-Hispanic white women, which has been attributed to both genetic and environmental risk factors, such as breastfeeding and diet (Fejerman & Ziv 2008). We aim to address this by examining the performance of the AI across different demographic groups, and by comparing its performance in the group of residents in Sweden that was involved in the trials to its performance where it was developed.

2.2.3 Local context

AI implementation is about fitting a new technology to the pre-existing organizational framework in terms of people, processes and technologies (Coiera 2019). Coiera (2019) emphasized that organizational networks may change over time as related to the introduction of a new technology. In the trial, the decision to introduce the AI to the first screening phase was a natural choice because this way it would unquestionably make independent assessments, and it is expected that this approach would fit well in the pre-existing organization.

2.3 AI-Health ethics

Ethical concerns related to AI in health care can be *epistemic*, related to inconclusive or misguided evidence; *normative*, such as transformative effects and unfair outcomes; and *overarching*, for instance pertaining to traceability and moral responsibility (Morley et al. 2020). At the institutional level, another ethical concern related to traceability regards a lack of clarity of liability, which could halt adoption (Morley et al. 2020). In the current trial, this is accounted for as the radiologists have the same role and responsibilities as they had in the standard setting. Specifically, there are two ethical concerns that we consider to be of particular relevance in this context:

2.3.1 Individual safety and trust

From the perspective of the individual, an important epistemic concern related to AI-Health includes misdiagnosis or missed diagnosis (Morley et al. 2020). In the trial, it is expected that the AI will increase sensitivity of the screening process, as it will be introduced in the form of a third independent radiologist in the first screening stage. Consequently, it is not expected to present any new risks regarding patient safety as compared to the standard setting. On the other hand, the AI may reduce specificity if it would generate excessive risk scores that the radiologists would rely on in the consensus discussion. This would result in more women being recalled for further diagnostics without good reason. This will be examined thoroughly in the trial.

2.3.2 Clinician integrity

Another ethical aspect of introducing AI-Health solutions relates to the deskilling of health care practitioners and a potential overreliance on AI-tools (Morley et al. 2020). To address this, the AI will be introduced as an independent reader in the initial screening phase and so it will not affect the radiologists' role at this stage. Nevertheless, in the consensus discussion there may be a risk for biases associated with human-AI interactions, such as *automation bias*, i.e., over-confidence in computer-generated solutions that results in the overriding of correct human decisions (Cummings 2004), or its contrast, *algorithm aversion*, i.e., lower confidence in an algorithm than a human despite higher accuracy of the former (Dietvorst et al. 2014). In an assessment of the latter, Dietvorst et al. (2014) found that people were less confident in algorithms that had made mistakes as compared to human forecasters who had made the same types of errors. In our analysis, we will therefore examine whether the radiologists' confidence in the AI-system increases or decreases over time. Further, we plan to investigate whether the readers tend to rely more, or less, on the AI technology when they are stressed.

Moreover, we will assess whether their trust in the technology will be influenced if the AI makes an odd assessment that lacks common sense, i.e., an evaluation that a radiologist would be unlikely to do, such as classifying a mamilla as an anomaly. Common sense is the ability to draw inference on the basis of fundamental knowledge about the world, and, as addressed in Shanahan et al. (2020), the lack of common sense has been a major challenge for AI since the field's beginning (McCarthy 1959) – which remains unsolved

(Davis & Marcus 2015). AI in the form of deep learning algorithms may have limited reliability in assessments of edge cases that are rare in the training data (Hamon et al. 2020), because such programs depend on large sets of training data to achieve high accuracy (LeCun et al. 2015), while a human may make good assessments in rare cases based on common sense only. If an AI fails to correctly handle such cases, it may fundamentally change the health care practitioners' perception of the system as reliable, and this may have important implications for subsequent adoption.

2.4 An early finding in the trial

An aspect that concerns both AI implementation and AI ethics is the fact that there may be a need for local tuning of an AI before deployment in a trial, which needs to fit well with the pre-existing organization and ensure patient safety. This relates to how key concepts within AI ethics, *accuracy* and *trust*, play out in clinical settings, and it may influence how the technology is perceived and adopted. In the current trial, this is illustrated by the following: the AI has been tuned based on historical data to detect the same number of cancers in the AI and radiologist combination as the historic combination of two radiologists; hence, it has not been tuned to function on its own as well as one radiologist. Preliminary findings indicate that this results in two phenomena:

First, on its own, the AI's sensitivity is generally lower than a radiologist's: it flags about 25-30% of the cases that a radiologist would do. Therefore, the practitioners may perceive that using the AI would not be good since it misses too many malign cases. To some extent, this is because the radiologists flag image findings and women who say they have a lump, while the AI evaluates only the image; among those who say they have a lump, relatively few do have cancer as compared to those with suspect image findings. Thus, it is important to demonstrate to the practitioners that the total number of cancers flagged by the AI and one radiologist is the same as the number flagged by two radiologists, otherwise their perception may be that the AI is flawed.

Second, the number of cases that continue to the consensus discussion almost doubles, since the AI and a radiologist tend to make more diverse types of false positive errors than two radiologists make, and most of the flagged cases are not cancer. To maintain trust, it is vital to explain that the increase in consensus discussions is likely to be compensated by a decrease in time spent on initial screening if the first stage involves one instead of two radiologists (in a later implementation). Hence, the perception that the AI has low sensitivity in the trial may influence the practitioners' confidence in it. We plan to examine this during the trial.

3 SUMMARY AND CONCLUSIONS

Health care is a setting where the need for innovative solutions is urgent, but where accuracy, patient safety and fair treatment are immensely important. This study presents our approach to analyzing human-computer interactions (HCIs) as an AI is deployed in breast cancer screening. Specialist physicians have trained for many years to review mammograms, and it is critical to create an understanding of the complex HCIs that are expected to take place in this setting. Two strengths of the trial include the complexity of the interaction as the AI is used to support the centerpiece skill of highly trained medical professionals, and the substantial number of interactions that will be included, which will generate longitudinal data of whether HCIs change over time.

We have outlined some important sociotechnical uncertainties in this clinical setting; for instance, one aspect pertains to how the uptake of the AI will evolve during the course of the trial, and another concerns how the radiologists will relate to a potentially odd assessment by the AI. We will further examine the cognitive biases of *automation bias* and *algorithmic aversion*. Key ethical aspects include ensuring patient safety and clinician integrity during the trial. For AI to benefit a lot of people in health care, many more trials like this one are imperative – while keeping the safety of the patients and the integrity of the practitioners intact.

ACKNOWLEDGMENTS

This research was partially funded by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation.

REFERENCES

Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., … & Guindy, M. (2019). Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology, 292(2), 331-342. DOI: https://doi.org/10.1148/radiol.2019182622

Coiera, E. (2019). The last mile: where artificial intelligence meets reality. Journal of medical Internet research, 21(11), e16323. DOI: https://doi.org/10.2196/16323

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. In AIAA 1st Intelligent Systems Technical Conference (p. 6313). DOI: https://doi.org/10.2514/6.2004-6313

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM, 58(9), 92-103. DOI: https://doi.org/10.1145/2701413

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114. DOI: https://doi.org/10.1037/xge0000033

Duffy, S. W., Tabár, L., Yen, A. M. F., Dean, P. B., Smith, R. A., Jonsson, H., ... & Chen, T. H. H. (2020). Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. Cancer, 126(13), 2971-2979. DOI: https://doi.org/10.1002/cncr.32859

Engström, E., & Strimling, P. (2020). Deep learning diffusion by infusion into preexisting technologies–Implications for users and society at large. Technology in Society, 63, 101396. DOI: https://doi.org/10.1016/j.techsoc.2020.101396

Fejerman L. & Ziv E. Population differences in breast cancer severity. Pharmacogenomics. 2008 Mar;9(3):323-33. DOI: doi: 10.2217/14622416.9.3.323. PMID: 18303968. DOI: https://doi.org/10.2217/14622416.9.3.323

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707. DOI: https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. (2019). AI opportunities for health care must not be wasted. Health Management, 19.

Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and explainability of artificial intelligence. Publications Office of the European Union. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC119336/dpad_report.pdf [Accessed Feb. 25 2021]

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500-510. DOI: https://doi.org/10.1038/s41568-018-0016-5

Johannessen, L. K., & Ellingsen, G. (2009). Integration and generification—agile software development in the health care market. Computer Supported Cooperative Work (CSCW), 18(5-6), 607. DOI: https://doi.org/10.1007/s10606-009-9097-8

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. DOI: https://doi.org/10.1038/nature14539

McCarthy J. (1959). Programs with common sense. in: Proceedings of the Teddington Conference on the Mechanization of Thought Processes. Her Majesty's Stationary Office. 75-91

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 113172. DOI: https://doi.org/10.1016/j.socscimed.2020.113172

Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., ... & Strand, F. (2020). External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA oncology, 6(10), 1581-1588. DOI: https://doi.org/10.1001/jamaoncol.2020.3321

Shanahan, M., Crosby, M., Beyret, B., & Cheke, L. (2020). Artificial intelligence and the common sense of animals. Trends in cognitive sciences. DOI: https://doi.org/10.1016/j.tics.2020.09.002

Spielkamp, M. (2019). Automating Society: Taking Stock of Automated Decision-Making in the EU. Bertelsmann Stiftung Studies 2019.