Designing for situated AI-human decision making: Lessons learned from a primary care deployment

Ben Wilson[1], Darren Scott[2], Matt Roach[1], Emily Nielsen[1] and Berndt Müller[1]

[1]*Swansea University, Swansea, UK*
[2]*Cardiff University, Cardiff, UK*

Abstract

We present a case study of AI deployment in a UK primary care (family doctor) setting. This demonstrates some of the challenges of real-world deployment of AI-human systems for decision-making. We use the seven domains of the NASSS (nonadoption, abandonment, scale-up, spread, and sustainability) framework to structure the presentation of our evaluation. We highlight three key lessons that should inform not only future deployments and evaluations, but future design work itself. The lessons are to attend to wider impacts, incorporate quality improvement and quality assurance techniques and employ participatory design, iterative development and formative evaluation.

Keywords

human-AI, decision-making, real-world evaluation, situated use, sociotechnical systems

Acknowledgements

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor HaDEA can be held responsible for them. Grant Agreement no. 101120763 - TANGO. For the purpose of Open Access, the author has applied a CC BY license to any Author Accepted Manuscript (AAM) version arising from this submission.

1. Introduction

Medical and scientific communities collectively display a lot of optimism around the potential of AI in the healthcare space [1]. In 2016, AI healthcare research attracted more funding than any other AI application area [2]. AI has the potential to positively support healthcare professionals' work, for example, through medical imaging [3]. A specific task application for AI in healthcare is that of decision support, where impressive AI performance is cited as

Proceedings of the 1st International Workshop on Designing and Building Hybrid Human–AI Systems (SYNERGY 2024), Arenzano (Genoa), Italy, June 03, 2024.
b.j.m.wilson@swansea.ac.uk (B. Wilson); ScottD15@cardiff.ac.uk (D. Scott); m.j.roach@swansea.ac.uk (M. Roach); e.e.nielsen@swansea.ac.uk (E. Nielsen); berndt.muller@swansea.ac.uk (B. Müller)
0009-0004-5663-5854 (B. Wilson); 0000-0001-7738-8662 (D. Scott); 0000-0002-1486-5537 (M. Roach); 0000-0003-2389-541X (E. Nielsen); 0000-0002-9590-4517 (B. Müller)

evidence that technological developments will improve the quality and speed of decisions and enhance the ability of professionals to make consequential decisions on care [4].

However, optimism within the research community is tempered by many questions. There are significant uncertainties around how AI systems work with human processes [5] in turn leading to questions about the efficacy of AI deployed in real contexts [6, 7, 8], as well as the generalisability of headline performance claims [9]. Further concerns arise around the safety of AI system deployments especially in relation to those exposed to the consequences of algorithm-influenced decisions [10, 11] and the ethical issues that arise when decisions are made with AI support [12] as well as the longer-term impacts of patterns of reliance that might develop [13]. Numerous frameworks have arisen in attempts to guide the ever-changing realm of AI design [14], but the rapidity of innovation in AI and its capabilities leaves not only the public but deploying organisations themselves as well as regulators struggling to keep up with the pace of change [15].

We present a number of lessons on the deployment of AI within situated healthcare contexts, with considerations on how better to approach the design and evaluation of such systems in future in order to ensure maximisation of safety and effectiveness. The lessons arise from a case study conducted as part of an NHS-funded project evaluating AI technologies to be deployed in the healthcare setting.

2. Case study: presented via the NASSS framework

The case study presented in this paper looked at a decision support system that utilises AI to support the initial request triage process in UK primary care. In the UK, the local organisational unit is known as a medical practice (or simply a practice). The steps from initiation of a request to a definitive encounter with one or more healthcare professionals is known as a pathway.

The evaluation team brought to the project a combination of expertise - AI research, Quality Improvement, Health Technology Assessment and User Experience research. Our approach retrospectively reflects the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [16]. NASSS resonates with our experience in observing healthcare organisations grappling with the real-world challenges of deploying technology. It emphasises the sociotechnical character of system and process change in healthcare and aims to address the prevalent tendency to place focus on technical innovation and de-contextualised performance. It does this by foregrounding how technology has interdependencies with human processes that affect what can become effective, sustainable improvements.

The case study is a contribution to the literature on seeking to improve human-machine synergy as there are few examples of real-world implementation in healthcare.

The following sections are organised under the seven domains used in the NASSS framework. The scope for applying the NASSS framework in full was limited by the terms of the evaluation. We had limited influence on the evaluation outputs, for example, since these were framed by the NHS body commissioning our involvement. However, we review all elements of the framework here - and their interaction - although we adapt the order in which the elements are usually addressed.

2.1. The Technology

The system developer created a clinically-informed primary care expert system to produce pathway recommendations. These recommendations are based upon a self-selected and self-completed online questionnaire chosen from around 100 distinct templates designed by the developer to cover the range of anticipated complaints (i.e., clinical conditions). Each questionnaire is completed by the patient themselves (or their relative, guardian or carer) as a request to the primary care practice for an appointment or advice.

The expert system consisted of a large body of clinically-informed rules triggered by key informational features appearing in the completed questionnaires. A proprietary NLP library was used to process all free text content - the objective being to identify text that would indicate a need to 'up-triage' (ie, raise the acuity of a triage suggestion). Free text was compared to an index of clinically concerning phrases using an experimentally determined fuzzy matching threshold. Detection in this module would escalate a response that would otherwise be determined by the rules engine acting on highly structured questionnaire responses.

An issue with expert systems can be that they require large amounts of expert labour to create. And because they are rule-based, they have a tendency to become brittle as they grow, so can be labour-intensive to maintain, too.

In the first iteration of the expert system, the pathway recommendation provided was limited to naming the *Team* to be assigned. That is, whether the case should be handled by a member of the physician team, someone from a different clinical team, an allied health professional, an individual from the community team or one of the administrative staff.

The developer learnt several lessons from this first iteration. There was a practical barrier to efficient usage in that practices needed a workflow that allowed them to review and interact with requests both before and after the triage decision. The dynamic nature of primary care workloads also meant that they needed to be able to monitor and respond flexibly to hour-by-hour changes in demand (as well as capacity, which was not always predictable) over the course of the working week. From the small set of engaged practices in their early user-studies, the developer learned both that they needed to provide a dedicated workflow (centred on a new feature they called a smart inbox) and that there was potential to do more than had been anticipated.

From these early user studies along with the maturation of the rules engine and questionnaire templates, the developer realised that suggestions could include not just the *Team*, but also an indication of the mode of interaction to be used *Mode* - for example, there could be a face-to-face appointment, a video consultation, a phone call or a text-based response - as well as an indication of how urgently the case should be addressed (*Urgency*). As a result, work was undertaken to add functionality. A later iteration of the system - and the one we evaluated - therefore augmented the *Team* suggestion with *Mode* and *Urgency* suggestions (see Fig 1 for an outline of the deployed system). However, development timelines and deployment commitments meant that the new features saw limited prototyping and user-testing, an outcome foreshadowed by warnings from Greenhalgh et al [16, p11]. Iteration requires repeated rounds of work on development, participation and evaluation to avoid preventable problems with implementation.

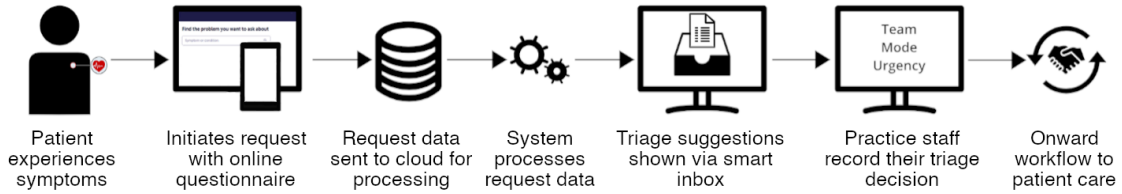

Figure 1: Triage suggestions system

2.2. The Adopter System (staff, patient, caregivers)

An important characteristic of the deployed system is that there are a multitude of people interacting with the workflow that might be considered users. The requestor (usually a patient) initiates a request to feed into the system by choosing which template to complete, using online guidance. The completed request is presented to primary care staff (a second user) in an on-screen workflow in which the three fields - Team, Mode, Urgency - can be selected for each case. When the AI suggestion feature is activated, AI recommendations for these three fields can be reviewed and then can be approved with a single click. Alternative selections can be made for each individual field if the user doesn't want to accept the recommendation. Following this triage process, there is an administrative step (which may be carried out by a different user) in which the triage decision is enacted as a booking that leads to the actual response event or encounter. Such booking steps often involve contacting the requestor by phone or text-based means to negotiate the details of a mutually suitable encounter (a possible third user). Following negotiation, there is always an appointing step within a calendar or diary system of some sort (a possible fourth user). The action or encounter thus made as a future commitment is ultimately fulfilled by a (potential fifth) user from whichever team has been assigned. At this point, the first user (the requestor, or patient) re-enters the orbit of the system as someone affected by the (AI-assisted) decision.

This wider view of who would be best engaged in a user study arises from looking at situated use. And even then, the variation in user configurations sometimes only becomes clear as new deployment sites are encountered. Foregrounding the sociotechnical character of implementation means anticipating that each deployment is like a new iteration, potentially requiring a further round of participation, development and evaluation.

Indeed, we found that the different staff involved in the workflow and their involvement with the system varied considerably by deployment site. Their response to deployment was also very varied. At some sites, engagement was very high. In others, it was extremely low. Our evaluation visits struggled to find relevant observable activity in some practices because the new workflow features were not part of established business-as-usual. At the same time, there were considerable complications related to patient participation. In some practices, a proportion of the patient body was reported to be reluctant to use the online request process and needed strong encouragement. In others, there was a lot of willingness on the part of the patients to submit online requests. However, the limited capacity of the practice to manage the resulting influx led to an approach to demand-management that involved throttling the access (closing down the online portal after a relatively short period or when a given number of requests had

been submitted). There was an assumption from the developer that workflow changes they had identified as improvements or enablers were unproblematic for each site to adopt. However, the diversity of sites meant that adoption entailed too much friction for too many practices. And design proposals to allow customisation to individual sites were necessarily delayed in favour of core functionality development.

2.3. The Organisation

An assumption of the NASSS framework is that adoption is considered in a single organisation and this is considered as a factor in its own right. The deployment we observed was, in fact, undertaken across several organisations at once. In an extension of the NASSS framework, we consider both the individual organisations and the variability between them as factors.

We identified a variability in practices' capacity to embrace change. This was certainly impacted by the challenging timelines imposed as a result of development delays. However, it also appeared to result from the assumptions from the developer concerning how workflows were managed. The practices most engaged with the early development of the system had high volumes of requests each day. To meet the demands of triaging large numbers quickly, they set up a dedicated and centralised team of medics carrying out triage. These triage teams worked on a rotation, so they were, at different times, both creators and consumers of triage decisions. In other practices, the use of online requests constituted only a very small fraction of the caseload and fewer resources were involved in handling the electronic triage process.

An even bigger factor for implementation was that some practices used non-medical staff with clinical training to execute the triage decision. And some made use of non-clinical (administrative) staff with clinical oversight and support. The effect of these significant variations was that the organisation of the workflow was different for each variant, and this was not fully anticipated in the development of the user-experience design. Important work was done to ensure a rapid workflow for clinical users who focused on this one triage decision task, which paid off for those practices whose processes aligned with the software. Other practices, with different workflows, had different needs from the technology. A greater understanding of the variety of workflows, and consideration for these differences in the design, may have led to greater adoption and uptake. We were less successful in finding out details of the adoption decision within each practice (that is, who made the decision to adopt). As in a lot of healthcare organisations, time is precious and disruption by researchers has to be kept to a minimum. We also ran into a seasonal period where 'organisational slack' was at a premium. Deployment that had been slated for the Spring of 2023 took place in the late Autumn and pushed evaluation into the busy Winter months.

Another component referred to by Greenhalgh et al is that of technological change embrittling existing routines [16, p13]. This may happen as a result of reducing opportunities for collaborative dialogue when automation shifts both the pace and context of a workflow. Our evaluation did not detect such reductions, but suggested that longer-term monitoring would be appropriate to ensure they weren't an unintended outcome.

One of the things we did identify was that deployment produced some notable changes not directly related to the algorithm or its intended effect. The enabling changes (to terminology as well as workflows) that were required to pave the way for the algorithmic contribution -

allowing the AI suggestions to have a feasible role within the human task, produced measurable changes in the triage process. An example was the shift from time-based urgency labels (eg, 24-hour, 3-day or 2-week appointment urgencies) to description-based urgency labels (urgent, soon, routine). These were implemented with a specific mapping between the eight old and the three new labels so that one would have expected an approximate match in the proportions of cases directed to equivalent-urgency slots. However, without any algorithmic influence, the proportions shifted significantly in response to the terminology change. This was a strong reminder that changing a workflow can alter outcomes - even if there's no algorithm deployment. Another consequence of the enabling changes required for the AI system was that downstream work to make appointments was altered. We weren't able to study the consequences of this in the scope of the evaluation. However, corrections for such alterations fit into the category of the hidden work of implementation (coherence work) mentioned by Greenhalgh et al [16, p13]. This change to downstream work has no impact on the operation of the AI from a metrics standpoint, but has clear impacts on the deployment space which may have a negative impact on overall efficacy of the system.

2.4. The Condition

As with most primary and many secondary care systems having a workflow functionality, all cases are potential candidates for processing irrespective of the clinical condition involved. This means many diverse conditions will be subject to the system with better or worse results. Where the workflow is manually streamed, the result is a subset being selected for inclusion or exclusion in a given process. However, the starting point (a worklist or 'inbox') is an artefact of the system. And with technology, this is frequently a given property fixed by designed. This is what the human must start from.

Using an electronic questionnaire or template means practice staff can filter the incoming requests on the basis of which template was completed. For primary care, there are a lot of encounters that are reviews of long term conditions or of previously prescribed medications. These, by nature, do not represent acute need. And they most often have a natural team assignment that follows the pattern of previous encounters. Other templates are used by patients to respond online to requests by the clinical team for information. The triaging software being deployed is designed to ignore these types of incoming submissions. However, the NLP-based escalation process can act on any request. The result of this is that some submissions get a suggestion that would normally not be processed by the suggestion algorithm. At the same time, if there is insufficient clear data to drive the rules engine on expected submissions, the system will not make any recommendation at all. So the cases getting a suggestion don't match up with the documented list of templates that would be expected to produce them.

For the system in question, one consequence of this departure from the condition-specific framework anticipated by NASSS is that complicating factors are not limited to comorbidities that alter the clinical features of a case. The technology we evaluated is not aimed at supporting particular needs associated with heart failure or the challenges associated with cognitive impairment, as in the single condition scenario. Cases all share the feature of being new presentations requiring a decision on management - in turn involving a greater or lesser level of diagnostic evaluation. The generality of the casemix means that the complicating factors

include simple-to-define issues, such as whether the requestor is submitting on their own behalf or for a relative or dependent. However, they also include highly complex human factors, such as whether a request on an 'abdominal pain' template is revealing significant information about a condition normally catered for by a different template, such as one for genital or reproductive system complaints. Or whether the information content reveals that the real issue is administrative, but it is recorded on a clinical template or *vice versa*.

Complexity here, then, arises not from the condition and its potential comorbidities, but from the fact that the system's main function is aiming to be condition-agnostic. The triage suggestion should be as effective for an ear infection as for an in-growing toenail. In order to cater for all possible submissions, the system has to cater for a huge variety of presenting detail and incorporate a large number of clinical caveats.

2.5. The Value Proposition

The value anticipated from deployment is assumed to be in the form of benefit for the primary care practice and its patients. If practice staff can complete triage more rapidly, it releases staff time for other work. And that benefits patients and allows more throughput. If triage is completed more accurately, then patients see the right team, with the right kind of encounter and with an appropriate urgency - again benefitting both the practice and its patients.

Our evaluation was necessarily limited to comparing before-and-after data on average triage times (using time and motion studies), comparing patterns of triage decisions (using process-behaviour analysis) and exploring the views of staff (using thematic analysis of interview records). While our evaluation suggested there were potential efficiency gains depending on the pre-existing workflow, there was no evidence of a change in accuracy. However, there was anecdotal evidence that there were costs to some patients in the additional effort required to secure an appointment request. And this needed further exploration.

Assuming the potential can be realised by further iterations of the workflow and tuning of the decision parameters, there is still the question of who benefits from the deployment. A challenge with technology deployments in many domains is determining whether releasing staff time really does improve the service. Further complexity is added in sociotechnical systems where staff recruitment and retention is difficult. In such contexts, technology deployments can help compensate for staffing challenges and ameliorate the negative effects of staff shortages or over-work. This is an overall improvement. There remains a question, however. Is the technology enabling better care or camouflaging more deeply rooted problems of sustainability? Patients who experience an overall improvement in access times (or a sustaining of access times by fewer staff) may not notice a drop in the quality of clinical pathway decisions. As with many clinical AI deployments, our evaluation was not able to access definitive ground truth values for prospective data.

A distinct challenge arises from incorporating an AI contribution into what is essentially a human-driven workflow. The system developer was pleased to be able to say that their clinically-specified rules engine avoided the governance issues and lack of transparency associated with deploying a 'black box' algorithm. Each decision could be audited in this deterministic system by replicating the clinical history as input and inspecting the response of each rule in the logic pathway to discover why a given suggestion had been made. But, as with all hybrid

decision-making, there is a critical difference between the facility to provide an explanation after-the-fact in the event that something was deemed to have gone wrong (eg, an audit process that allows interrogation of the reason for an aberrant suggestion following an incident) and any transparency that forms a part of the normal workflow (eg, an explanation of the algorithm's contribution to the human decision-maker at run-time). In other words, the system's interpretability could not contribute synergistically to the decision itself.

2.6. The Wider Context

In UK primary care, the context is that the government, immediately prior to the COVID-19 pandemic, was heavily pressing for increased use of technology, both to organise consultations and to replace face-to-face appointments. The pandemic period saw this process massively accelerate with increased use of online requests and remote appointments. However, following the pandemic period, a surge in pent-up demand coincided with a government campaign to support face-to-face appointments in a reversal of the previous policy direction. At the same time, there is a lot of political pressure on primary care to increase productivity, care for more patients and sustain services in conditions of real-terms funding reductions. And there is a wider UK health service context of immense pressure on hospital emergency departments and associated services. Technology is seen as having the potential to alleviate capacity pressures without increasing staffing costs. However, extreme pressures can all too easily lead to reliance on technology which does not account for complexity in how clinicians, staff and patients work together, especially if the operating patterns differ between sites and between individual clinicians.

2.7. Interaction over Time

Our evaluation was time-limited and hence not able to review the medium- and long-term feasibility of continuing to adapt to the technology. In addition to the ongoing sociotechnical adaptations influenced by the deployed technology, there is a purely technical challenge in maintenance. In medicine, updates to clinical knowledge and guidelines as well as perspectives on best practice are made constantly. Incorporating these ongoing updates is a constant source of work. However, it is made considerably bigger in complex software with the growth of the testing burden. We witnessed the beginning of this during the project. Before full deployment, there had been an update to some questionnaire templates and these new designs had not been incorporated into the algorithmic decision support owing to the development and test burden.

A separate issue that requires a longer timeframe to explore is the alteration to the skill-mix of staff. This can apply to both individual staff where decision support leads to the atrophying of cognitive skills - and to bodies of staff where recruitment does not secure certain skillsets because they are no-longer considered essential. In each case, the human component of the synergistic system is altered and its role in the combination is likely to change. Such changes may or may not represent a risk to the decision process as a whole [17]. However, even where the decision process is protected, there is a need for consideration of the human impact of such de-skilling.

3. Lessons Learnt

3.1. Lesson 1: Wider Impacts of AI Deployments

As with any technological change in a sociotechnical system, there needs to be a recognition that change not only impacts the immediate context, but also many of the surrounding processes. Change creates ripples in the wider situated space which must be accounted for when evaluating the effectiveness and safety of the technology.

Within the example of triage, as we saw above, the downstream effects of changed pathways or terminologies might be to alter workloads - whether for clinical roles or administrative functions. If this change in workload is a result of a less appropriate decision, then the change cannot be regarded as an improvement - regardless of any apparent efficiency gain.

Holistic evaluation requires us to discuss not only the technology and its inherent value, but the ability of the adopter and the organisation more generally to accept the technology and effectively integrate it. The interplay between the technology and its adopter, between the actions of the system and the responses of those linked to it, creates consequences for different actors in the system. These consequences are what we discuss here as 'ripples'.

This demonstrates the necessity of a situated evaluation - these ripples can only be observed in a context that is (or approaches) a real-world deployment. By approaching the evaluation of AI with participation of stakeholders and early rounds of iteration in a situated context, we can best ensure the deployment of AI is providing the intended benefits rather than creating additional issues.

3.2. Lesson 2: Quality Improvement & Assurance

Change in sociotechnical systems always requires careful evaluation to ensure that change is an improvement rather than change for its own sake. And to ensure appropriate focus is maintained on the target-for-change while monitoring associated processes for any sign of unintentional change. These approaches must be underpinned by assurance processes that provide confidence that measures are meaningful and accurate.

Our evaluation included both time and motion elements and semi-structured interviews with practice staff as well as standard algorithm performance metrics. This ensured that the evaluation was not exclusively focused on a narrow set of technical measures but enabled a rounded picture of the effect of introducing both the enabling changes and the algorithm itself.

Ensuring that the techniques of process improvement, quality improvement and quality assurance are central to the evaluation of AI allows us to move past the static 'product acceptance' model to something more agile, allowing for continuous adoption and evaluation.

3.3. Lesson 3: Participatory Design, Iterative Development and Formative Evaluation

As is evident from the two lessons above, the dominant model of AI evaluation is not well suited to healthcare due to the sociotechnical nature of healthcare organisations and the complex nature of healthcare decision-making. We argue for changes to the process of evaluating AI

in the critical healthcare space, but we also go beyond evaluation to argue for fundamental changes to the design of algorithms for the healthcare space.

To account for unforeseen issues in deployment, the development of AI for healthcare contexts must adopt a more participatory approach. While the development of AI by experts pursuing well-formulated problem statements with clear ground truths and well-defined objective functions has its obvious merits, it can miss important nuances of the target deployment space which, if incorporated early can allow for greater effectiveness and more valuable algorithmic contributions. Healthcare AI requires greater input and engagement from healthcare professionals, patients and wider stakeholders to ensure both that the proposed algorithm is beneficial (or non-detrimental) to all parties, and that the design of the system is able to realise these benefits when deployed into the target context.

The design of AI for healthcare must also adopt a more iterative approach, including additional design cycles and formative evaluation processes to ensure the system is developed in line with real-world requirements and is ultimately suitable for its purpose and its situated context. The resulting fairness and transparency of the human-algorithm system must be considered as much a part of the development and evaluation process as the required performance metrics.

4. Conclusions

As AI becomes more ubiquitous within the healthcare space, it should become increasingly clear that existing processes of AI development and deployment themselves need to be improved. From our work, we have surfaced three key lessons to be carried forward into the design and evaluation of healthcare AI. Ensure the wider impacts of technology adoption are captured as part of the evaluation. Apply high standards of quality improvement to ensure overall change is actually beneficial (measure the right things) while applying equally high standards of quality assurance to ensure the various measures are reliable (measure them properly). We also promote a change in the overall culture of AI design within this sector - government, healthcare, and technology leaders need to embrace not only the potential of AI itself, but careful consideration of situated use alongside attention to algorithm fairness and transparency to assure technology deployments are able to best benefit patients and clinical staff rather than introduce additional issues. It is vital to assign accountability across developers, clinicians, commissioning health organisations, patients and regulatory authorities to ensure inevitable concerns are properly addressed. As the AI creates ripples, all those in contact with these ripples have a part to play in ensuring all technology deployments are safe and effective, able to best enhance the quality of healthcare for all parties.

When AI systems are developed well and deployed effectively they can revolutionise service delivery, improve patients' lived experience and even save lives. Our experiences have shown us that this quality of development and deployment is still an ideal to be achieved. The rapid progress in technical development of algorithmic systems needs to be matched by progress in directing those developments towards demonstrated situated benefit.

References

[1] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthcare Journal 6 (2019) 94–98. URL: /pmc/articles/PMC6616181/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/. doi:10.7861/futurehosp.6-2-94.

[2] V. H. Buch, I. Ahmed, M. Maruthappu, Artificial intelligence in medicine: Current trends and future possibilities, 2018. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5819974/. doi:10.3399/bjgp18X695213.

[3] R. Han, J. N. Acosta, Z. Shakeri, J. P. Ioannidis, E. J. Topol, P. Rajpurkar, Randomized Controlled Trials Evaluating AI in Clinical Practice: A Scoping Evaluation, 2023. URL: http://medrxiv.org/lookup/doi/10.1101/2023.09.12.23295381. doi:10.1101/2023.09.12.23295381.

[4] S. L. Goldenberg, G. Nir, S. E. Salcudean, A new era: Artificial intelligence and machine learning in prostate cancer, Nature Reviews Urology 16 (2019) 391–403. URL: https://www.nature.com/articles/s41585-019-0193-3. doi:10.1038/s41585-019-0193-3.

[5] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, C. Tan, Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, ACM, Chicago IL USA, 2023, pp. 1369–1385. URL: https://dl.acm.org/doi/10.1145/3593013.3594087. doi:10.1145/3593013.3594087.

[6] F. Cabitza, J.-D. Zeitoun, The proof of the pudding: In praise of a culture of real-world validation for medical artificial intelligence, Annals of Translational Medicine 7 (2019) 161. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6526255/. doi:10.21037/atm.2019.04.07.

[7] F. Cabitza, A. Campagner, F. Del Zotti, N. Verona, A. Ravizza, F. Sternini, All you need is higher accuracy? On the quest for minimum acceptable accuracy for Medical Artificial Intelligence, in: Proceedings of the 12th IADIS International Conference e-Health 2020, EH 2020 - Part of the 14th Multi Conference on Computer Science and Information Systems, MCCSIS 2020, 2020, pp. 159–166.

[8] J. S. Marwaha, J. C. Kvedar, Crossing the chasm from model performance to clinical impact: The need to improve implementation and evaluation of AI, npj Digital Medicine 5 (2022) 1–2. URL: https://www.nature.com/articles/s41746-022-00572-2. doi:10.1038/s41746-022-00572-2.

[9] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, E. K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, PLoS Medicine 15 (2018). doi:10.1371/journal.pmed.1002683.

[10] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in ai safety, arXiv preprint arXiv:1606.06565 (2016). URL: http://arxiv.org/abs/1606.06565.

[11] S. Saisubramanian, S. Zilberstein, E. Kamar, Avoiding negative side effects due to incomplete knowledge of ai systems, AI Magazine 42 (2022) 62–71. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/7390. doi:10.1609/aaai.12028.

[12] R. V. Zicari, J. Brusseau, S. N. Blomberg, H. C. Christensen, M. Coffee, M. B. Ganapini, S. Gerke, T. K. Gilbert, E. Hickman, E. Hildt, S. Holm, U. Kühne, V. I. Madai,

W. Osika, A. Spezzatti, E. Schnebel, J. J. Tithi, D. Vetter, M. Westerlund, R. Wurth, J. Amann, V. Antun, V. Beretta, F. Bruneault, E. Campano, B. Düdder, A. Gallucci, E. Goffi, C. B. Haase, T. Hagendorff, P. Kringen, F. Möslein, D. Ottenheimer, M. Ozols, L. Palazzani, M. Petrin, K. Tafur, J. Tørresen, H. Volland, G. Kararigas, On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls, Frontiers in Human Dynamics 3 (Jul-2021). URL: https://www.frontiersin.org/articles/10.3389/fhumd.2021.673104.

[13] K. M. Kostick-Quenet, S. Gerke, AI in the hands of imperfect users, npj Digital Medicine 5 (2022) 1–6. URL: https://www.nature.com/articles/s41746-022-00737-z. doi:10.1038/s41746-022-00737-z.

[14] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. URL: https://www.nature.com/articles/s42256-019-0088-2. doi:10.1038/s42256-019-0088-2.

[15] K. Ganapathy, Artificial Intelligence and Healthcare Regulatory and Legal Concerns, Telehealth and Medicine Today 6 (2021). URL: https://telehealthandmedicinetoday.com/index.php/journal/article/view/252. doi:10.30953/tmt.v6.252.

[16] T. Greenhalgh, J. Wherton, C. Papoutsi, J. Lynch, G. Hughes, C. A'Court, S. Hinder, N. Fahy, R. Procter, S. Shaw, Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies, Journal of Medical Internet Research 19 (2017) e367. URL: https://www.jmir.org/2017/11/e367. doi:10.2196/jmir.8775.

[17] F. Cabitza, R. Rasoini, G. F. Gensini, Unintended Consequences of Machine Learning in Medicine, JAMA 318 (2017) 517–518. URL: https://doi.org/10.1001/jama.2017.7797. doi:10.1001/jama.2017.7797.