RESEARCH-ARTICLE

"It depends": Configuring AI to Improve Clinical Usefulness Across Contexts

HUBERT DARIUSZ ZAJAÇ, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

JORGE MIGUEL RIBEIRO, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

SILVIA INGALA, Rigshospitalet, Copenhagen, Hovedstaden, Denmark

SIMONA GENTILE, Rigshospitalet, Copenhagen, Hovedstaden, Denmark

RUTH WANJOHI, Nairobi Hospital, Nairobi, Nairobi, Kenya

SAMUEL NGUKU GITAU, The Aga Khan University Kenya, Nairobi, Nairobi, Kenya

View all

Open Access Support provided by:

Rigshospitalet

University of Copenhagen

The Aga Khan University Kenya

Nairobi Hospital

Published: 01 July 2024

Citation in BibTeX format

DIS '24: Designing Interactive Systems Conference
July 1 - 5, 2024
Copenhagen, Denmark

Conference Sponsors:
SIGCHI

"It depends": Configuring AI to Improve Clinical Usefulness Across Contexts

Hubert D. Zając
hdz@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

Jorge M. N. Ribeiro
jori@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

Silvia Ingala
silvia.ingala@regionh.dk
Rigshospitalet Copenhagen
University Hospital
Copenhagen, Denmark

Simona Gentile
simona.gentile@regionh.dk
Rigshospitalet Copenhagen
University Hospital
Copenhagen, Denmark

Ruth Wanjohi
ruthwanjohi@nbihosp.org
The Nairobi Hospital
Nairobi, Kenya

Samuel N. Gitau
samuel.gitau@aku.edu
Aga Khan University Hospital
Nairobi, Kenya

Jonathan F. Carlsen
jonathan.frederik.carlsen@regionh.dk
Rigshospitalet Copenhagen
University Hospital
Copenhagen, Denmark

Michael B. Nielsen
michael.bachmann.nielsen@regionh.dk
Rigshospitalet Copenhagen
University Hospital
Copenhagen, Denmark

Tariq O. Andersen
tariq@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

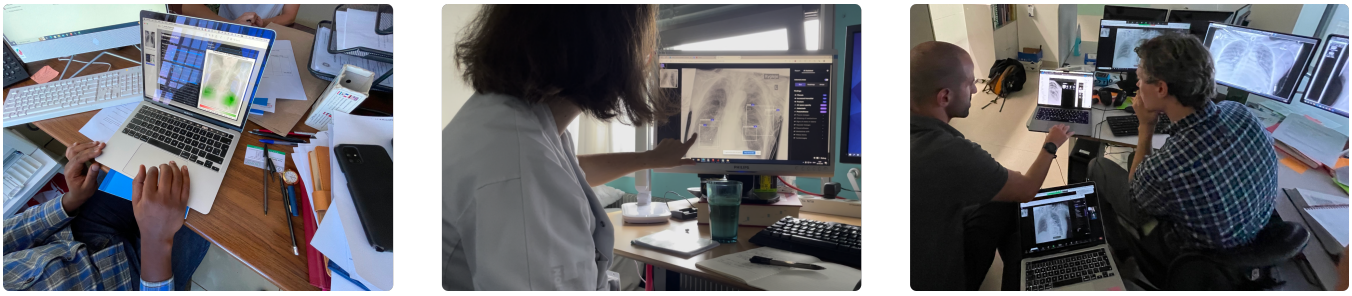

Figure 1: Design interventions in different radiology settings in Kenya and Denmark. Three versions of an AI-based prototype were used to explore configuration opportunities to achieve clinical usefulness across clinical sites (from left: version I, II, II).

ABSTRACT

Artificial Intelligence (AI) repeatedly match or outperform radiologists in lab experiments. However, real-world implementations of radiological AI-based systems are found to provide little to no clinical value. This paper explores how to design AI for clinical usefulness in different contexts. We conducted 19 design sessions and design interventions with 13 radiologists from 7 clinical sites in Denmark and Kenya, based on three iterations of a functional AI-based prototype. Ten sociotechnical dependencies were identified as crucial for the design of AI in radiology. We conceptualised four technical dimensions that must be configured to the intended clinical context of use: AI functionality, AI medical focus, AI decision threshold, and AI Explainability. We present four design recommendations on how to address dependencies pertaining to the medical knowledge, clinic type, user expertise level, patient context, and user situation that condition the configuration of these technical dimensions.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Machine Learning, Human-Centred AI, User-Centred Design, AI Interaction, Performance Optimisation, Usability, Transferability

ACM Reference Format:
Hubert D. Zając, Jorge M. N. Ribeiro, Silvia Ingala, Simona Gentile, Ruth Wanjohi, Samuel N. Gitau, Jonathan F. Carlsen, Michael B. Nielsen, and Tariq O. Andersen. 2024. "It depends": Configuring AI to Improve Clinical Usefulness Across Contexts. In *Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3643834.3660707

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3660707

1 INTRODUCTION

Artificial Intelligence (AI) models repeatedly match or outright outperform radiologists in narrowly defined detection tasks [4, 53, 60, 63]. There are multiple studies claiming that AI-based systems enhance radiologists' work, either by increasing accuracy or reducing time spent on each examination [48]. These claims, however, are based on retrospective evaluations conducted in laboratory settings. When looking closer into the state of the art of clinician-facing AI, the claims of utility weaken [93]. For example, Roberts et al. [61] found that out of the 62 AI models detecting and predicting COVID-19 on chest X-rays and CT scans that were described in the literature, none were deemed to be useful for clinical purposes. Furthermore, evaluations of the handful of systems approved by the authorities in the United States and European Union [1] revealed that their clinical impact when integrated into practice remains mostly unclear [79, 84]. A similar study by Lehman et al. [47] showed no improvement in patient outcomes after the successful integration of an AI-based support tool for mammography screenings. Strohm et al. claimed that one of the primary causes of AI's lack of success in radiology until now is due to "uncertain added value for clinical practice of AI applications" [73]. What these studies show is that the clinical usefulness of hitherto AI-based support systems is limited.

Researchers with diverse backgrounds (AI, Health, and Human-Computer Interaction (HCI)) investigated what makes AI-based support systems clinically useful. Based on the previous work, we define clinical usefulness as the overarching quality of AI-based support systems emerging from the interplay of their real-world performance, clinical efficacy, local applicability, and end-user acceptance in a situated clinical context for concrete end-users. First, robust performance in real-world settings is essential, as subpar performance has been found to increase workload and disrupt clinical routines [80, 82, 93]. Second, the evaluations, primarily assessing technical performance metrics through randomised clinical trials (RCTs), must encompass tangible clinical outcomes and patient benefits. Health researchers have been advocating for more flexible assessment methodologies aligned with the iterative nature of AI deployment [11, 42, 49]. Third, end-user acceptance, supported by qualities like trust and usability, emerges as pivotal for successful use in clinical practice [18, 19, 39]. Altogether, for an AI-based system to be clinically useful, it must perform well, benefit patients, and be accepted by clinical end-users working in different clinical contexts.

In this paper, we investigate **how to design AI for clinical usefulness in different clinical contexts.** This study was conducted as a part of a larger research and development project focused on innovating an AI-based system to assist examinations of chest X-rays in Denmark and Kenya. Here, we define innovation as the entirety of work conducted to create an AI-based system, from creating the datasets the AI is trained on through design and development to its integration and use in practice. We conducted 19 design sessions and design interventions (online and collocated) with 13 radiologists from 7 clinical sites in Denmark and Kenya. Throughout the design study, we explored a range of user interface mock-ups and three versions of a web-based prototype of an AI-based support system with prioritisation and decision-support functionalities.

We conceptualised four technical dimensions of radiological AI support that need to be configured to maximise its clinical usefulness. The technical dimensions uncovered through the design interventions span *AI functionality, AI medical focus, AI decision threshold, and AI Explainability*. These decisions constitute the critical aspects of radiological AI-based support systems and must be configured in relation to the local social dimensions of clinical AI.

Moreover, to support configuration during innovation, we deconstructed social dimensions, conditioning how each of the technical dimensions supports final clinical usefulness. Namely, how *medical knowledge, clinic type, user expertise level, patient context*, and *user situation* affect the clinical usefulness of the technical dimensions.

Finally, we discuss how these dependencies should be accounted for throughout the innovation processes to successfully configure future systems before-use and enable meaningful configuration in-use. Based on the design interventions, we offer four concrete design recommendations addressing the configuration needs of each of the conceptualised technical dimensions of clinical AI.

2 RELATED WORK

2.1 Clinical Usefulness of AI Systems in Healthcare

The hitherto evidence of AI's positive influence on clinical practice is limited [51, 75, 80]. Research on the real-world effect of AI in healthcare tends to be discrete and focusing on confined goals [93]. However, to provide clinical value AI-based systems have to dovetail contributions from Human-Computer Interaction, AI, and Health into a cohesive vision [30, 82, 93].

First, clinical usefulness necessitates robust performance [93]. This primarily has to be true in real-world settings, retrospective evaluations in lab environments do not speak to the final performance of a system. For example, in a real-world evaluation of an acclaimed ML model for detecting diabetic retinopathy, 21% of all cases were deemed ungradable [8]. Poor performance also leads to increased workload [57, 64, 82], additional time spent on discerning false positive predictions [67, 68], or breakages to work routines [31]. Van Leeuwen et al. [80] reported that out of 100 CE-approved radiological AI-based systems, 64 showed no peer-reviewed evidence of clinical efficacy. Most evidence for the remaining 36 systems focused on diagnostic accuracy, not real-world clinical outcomes.

Second, clinical usefulness necessitates clinical efficacy [42]. However, randomised clinical trial (RCT) - a focused, systematic, rigorous, and insulated method commonly used to evaluate the validity of clinical interventions independent of external confounders - is often following the traditional sequential paradigm of work characteristic for drug development [13]. In this tradition, the intervention is evaluated only when deemed complete [21]. When translating this mentality to AI-based systems, not only does it hinder innovation, but it also results in the evaluation of AI through the measure of technical performance [49]. While technical performance is the backbone of useful AI, clinical efficacy is not its immediate consequence [11, 69]. For example, Lehman et al. [47] conducted a prospective evaluation of a computer-aided detection system supporting mammography reporting. Researchers concluded that the

use of AI had no "established benefit to women." Instead, healthcare researchers are opening up towards more flexible evaluation approaches that align with the iterative and situated nature of AI innovation and "go beyond measures of technical accuracy to include quality of care and patient outcomes" [23, 42]. Achieving high performance but in metrics that are clinically relevant is the next step towards clinically useful AI-based systems.

Third, clinical usefulness necessitates clinical organisational acceptance. HCI community's claim to fame is understanding that regardless of a system's performance, it will not have any impact if no one wants to use it. Thus, many facets of making clinical AI an appealing solution were explored. Trust has been hallmarked as a critical quality of clinical AI. HCI researchers investigated its origin [5, 18] and dependencies [59], as well as issued recommendations for design [39]. Explainable AI (XAI) has been the most promising answer to enhance trust, support oversight, and increase the perceived usefulness of clinical AI [19, 34, 46, 86]. AI as a new source of information and agency prompted the exploration of new ways of reasoning and human-AI collaboration [10, 16, 20, 25]. Researchers also investigated AI's position in a clinical decision-making process [41] and the rationale behind integration opportunities into clinical practice [39, 66, 68, 90]. They argued that the workflows, current work practices, and the broader sociotechnical context should also be taken into account when implementing clinical AI-based systems [22, 39, 56, 62, 76, 93]. Addressing these concerns is crucial for AI to have a chance at benefiting patients and being accepted by healthcare professionals.

Altogether, for an AI-based system to be clinically useful it must perform well, benefit patients, and be accepted by clinical end-users. However, oftentimes the innovation of clinical AI is conducted in silos and the work is not guided by the ultimate goal of clinical usefulness [13, 93]. We need to investigate how AI-based systems can be configured to support these three goals and ultimately result in clinically useful AI.

2.2 System Configurability

Configurability has been long considered crucial to the appropriation of IT systems [26, 27, 45]. There are two types of configurability that should be explored in the context of this study: before-use and in-use [36].

Before-use configurability typically involves the active participation of end-users in the design processes, aiming to tailor systems to their specific needs and preferences [36]. Various methods and approaches have emerged to facilitate meaningful engagement with end-users, such as participatory design techniques [45]. Acquiring an understanding of work practices and work environment, but also technology aspects of a future system and changes it may introduce, is critical for developing systems that effectively respond to user needs [43]. This understanding enables developers and designers to implement systems that are not only technically sound but also contextually appropriate.

However, according to Stewart and Williams [72], the paradigm of user-centred design does not properly answer the challenges of implementing useful systems. Rather, the final usefulness of a system is created iteratively through the acts of in-use configuration. This stance echoes Suchman who recognised the need for design activities to continue after a system's deployment [74].

The in-use configuration may cover functionalities, user interface, or other settings that let the end-users adjust the system to their preference and work environment [85]. However, the system is not the only configurable arena. The environment also undergoes a process of configuration to the new system. The in-use configuration processes encompass changes to the "technical environment, organisational relations, space technology relations, as well as people's connections to other people, to other places, and work materials" [6]. Dourish highlights how the appropriation of IT systems in practice is an act of both adapting the technology and adapting the practices to fit into the new reality [27].

As usual with AI, the matter of configuration is burdened by the immutability of certain aspects of the system in-use and the dependency of early design decisions on the use context [93]. HCI researchers investigating the design of AI-based systems learned that it is impossible to envision all aspects of clinical AI-based systems before deployment. As a result, the final capabilities of such systems only take shape after they have been deployed. [33, 88]. On the opposite end of AI innovation, i.e., prior to data labelling, Zając & Avlona [93] established that very concrete choices and assumptions about the final context of AI use form the data used for AI training and, by extension, shape the space of capabilities of future AI-based systems. This vicious cycle of dependencies prompted researchers into new ways of thinking about AI innovation. Edwards et al. [28] proposed the concept of "growing" to foreground the need for almost organic adoption and adaptation of new IT systems in an existing environment. Elish and Watkins presented a similar argument [29] who emphasise that early realisation of clinical AI and acknowledgement and support of the necessary "repair work" are crucial to counter the risk of a system remaining "a potential solution", i.e., a solution that is not viable when actually implemented.

We see the problem of configuration of clinical AI, as a problem of obtaining reliable information related to design decisions made during the innovation process. The emergence and propagation of dependencies (or "sociotechnical interdependencies", see [93]) at the point of deployment hamper the ability to configure clinical AI-based systems in-use. At this point, the assumptions about the context of use are already ingrained in the AI model. We want to support the configuration of radiological AI-based systems for clinical usefulness by uncovering the dependencies anchored in clinical contexts and linking them with specific design decisions. This extended understanding of contextual factors will allow developers and designers to implement radiological AI support configurable and useful across clinical contexts.

3 METHODOLOGY

In this paper, we explored how to design radiological AI-based systems for clinical usefulness across contexts. This study was part of a larger project set to design and develop an AI-based support tool for radiologists examining chest X-rays, funded by the Innovation Fund Denmark (0176-00013B). The project is a multidisciplinary collaboration between the Department of Computer Science at the

#	Type	Radiologists (Total)	Country
K1	Small General Hospital	1 (<5)	Kenya
K2	Small General Hospital	1 (1)	Kenya
K3	Imaging / Teleradiology Clinic	1 (1)	Kenya
K4	Specialised Hospital	1 (<20)	Kenya
K5	Big General Hospital	3 (10)	Kenya
D1	Specialised Hospital	6 (>100)	Denmark
D2	Imaging Clinic	1 (<5)	Denmark

Table 1: Clinical sites included in the study and the number of radiologists (N=13) participating from each of the sites in relation to the total number of employed radiologists. One radiologist with double affiliation in D1 and D2.

#	Participant	Expertise	Clinical site	Length	Prototype
I01	P01	Senior	K1	60m	I
I02	P02	Senior	K2	60min	I
I03	P03	Senior	K3	120min	I
I04	P04	Senior	K4	50min	I
I05	P05	Senior	K5	80min	I
I06	P06	Senior	K5	80min	I
I07	P07	Senior	K5	60min	I
S08	P08	Junior	D1	70min	II
I09	P09	Junior	D1	50min	II
I10	P10	Senior	D1	30min	II
I11	P11	Senior	D1 / D2	30min	II
S12	P08	Junior	D1	60min	II
S13	P10	Senior	D1	30min	III
S14	P12	Junior	D1	80min	III
I15	P11	Senior	D1 / D2	45min	III
I16	P10	Senior	D1	30min	III
I17	P13	Junior	D1	40min	III
S18	P05	Senior	K5	95min	III
S19	P04	Senior	K4	70min	III

Table 2: Radiology participants that took part in the study. Interventions - I, Sessions - S.

University of Copenhagen, the Department of Radiology at Rigshospitalet Copenhagen University Hospital, and Unumed ApS. Due to the project's goals, the future system should support radiologists in Denmark and Kenya. To take into account the diversity of practices and contexts, we conducted design research in seven different healthcare settings across the two countries (Table 1), which included: (1) imaging clinics - where medical imaging services such as chest radiographs, ultrasounds, or CT scans (only K3) are provided to patients referred by external physicians, (2) general hospitals - that offer primary and secondary care and refer patients requiring more specialised care to other facilities, and (3) specialised hospitals - that provide tertiary and quaternary care, handling the most complex medical procedures in their respective countries.

The participants were recruited through email and the professional networks of the project members. Nine senior (consultant) radiologists and four junior (in-training) radiologists joined the study (Table 2). Junior radiologists' reports must be most often approved by a senior colleague before sharing with clinicians. The senior radiologist's assessment is final. Participants were not compensated, and we collected written consent from all participants. According to the authors' institutions' institutional review boards (IRBs), our study was considered non-interventional and thus exempt from a formal ethical review.

3.1 Research Through Design: Design Interventions with Working Prototypes

To explore the clinical usefulness of AI in different radiology contexts, we undertook a research through design approach [96]. We conducted three iterations based on a series of design sessions and design interventions using mock-ups of user interfaces (Prototype I) and working prototypes (II and III) (Fig. 1, Fig. 2, and Fig. 3). The three iterations were determined by decisions to deploy major changes in the web-based prototypes, i.e. version 1-3, followed by gathering feedback from the participants. The design sessions were carried out both online and collocated with radiologists in hospital offices. During these sessions, we obtained medical domain knowledge, typically by clarifying questions about radiology work and X-rays, but we also collectively explored the design space through a range of mock-ups and prototypes. The design interventions were carried in-situ with the performative purpose of exploring how the proposed solutions would be enacted close to real-world radiology practices. A design intervention, as defined by Halse and Boffi [35], is a method that integrates design and ethnography and "enables new forms of experience, dialogue, and awareness about the problem to emerge" (see also [14, 15]). It is an experimental form of inquiry that enables a positioning "in-between what is already there and what is emerging as a possible future" [3].

In our case, this meant that we intervened in the radiologists' everyday work settings with design artefacts as a vehicle for exploring the dependencies of AI usefulness in situated contexts. During observations of the radiologists' work practices, we brought in the prototypes and mock-ups as a way to enact while experimenting with new forms of AI support in radiology. The benefit of this approach was the possibility to engage radiologists in moving between considering the proposed solutions and envisioning alternatives while constrained by the requirements of the local context. This mode of research was important for this study because it provided more grounded and realistic visions of how AI could become clinically useful across hospital contexts.

In total, we conducted thirteen design interventions and six design sessions (Table 2) with thirteen radiologists in Denmark and Kenya, lasting between 30 and 120 min (avg. 60 minutes). In between sessions and interventions, we designed a range of user interface mock-ups using Figma, consisting of different AI functionalities and alternatives to interactive features. A total of three versions of a web-based prototype, which included an AI model developed in the greater part of the project. This meant that the participants in this study interacted with real data and real output from the AI model during design interventions. Importantly, the data was completely anonymised, and no other medical information about the patients was available. The mock-ups, prototypes and feedback from the participants became input for multiple design meetings within a group of three of the authors (HDZ, JMNR, TOA) and

Figure 2: Online and co-located design sessions with user interface mock-ups and functional prototypes (from left: iteration I, II, and III)

included high-level discussions with machine learning engineers at Unumed ApS. Here, insights were discussed, and decisions were made regarding what the following design explorations should consist of. All design sessions and interventions were audio-recorded and machine-transcribed to support thematic analysis.

3.1.1 Prototypes. As part of the greater project, a deep learning-based model was developed by machine learning engineers at Unumed ApS to detect selected radiological findings [92]. The AI model was developed using a convolutional neural network. The first prototype was merely a proof of concept, not designed to collect feedback from external domain experts. It was developed to guide future work in terms of model development and data labelling. However, inspired by earlier research [92, 93], we considered it an opportunity to engage in more concrete discussions on the merit of clinical usefulness with medical professionals at an early stage of the innovation work.

The second and third iteration of the prototype consisted of an interactive web application designed to emulate a DICOM viewer. The web application integrated with the AI model developed within the bigger project. This connection enabled us to work with real data and, thus, explore with fidelity the interactions of the radiologists with the system. For the design interventions, radiologists were given access to the prototype, either in-person or remotely. They were requested to choose the next examination to report, following their usual practice and using information displayed in the prototype. Then, they were asked to interpret the selected examination without the use of AI and with AI decision support. Moreover, they were asked to configure the AI tool using available options to fit their practice. Finally, they were encouraged to explore the prototype independently and interact with any element of the user interface.

3.2 Analysis Positionality

The data analysis was conducted by the first and last authors (HDZ and TOA) with backgrounds in Health Informatics, HCI, and AI (5+ & 15+ years of experience). Moreover, before the analysis of the data from the design sessions and design interventions, the two co-authors concluded extensive ethnographic investigations into the work practices of radiologists from the visited sites with a particular outlook on opportunities for AI support (described in a manuscript prepared for publication). First-hand experience with the work practices and similarities and differences across clinical settings informed the initial analysis of this data.

3.3 Data Analysis

We used reflective thematic analysis [17] to analyse collected data (transcriptions of the design interventions). The analysis took place in Dovetail - a web application for qualitative data analysis. Except for the transcription software, no AI-based analysis support was used in this study. The two authors familiarised themselves with the collected data after every iteration of the design sessions and design interventions when deciding on the next focus. Moreover, the two authors, prior to coding, based on their fieldwork experience (60+ hours) and a literature review [93], devised three bucket themes to support the later organisation of codes: type of clinical site, domain expertise of medical professionals, and patient and situational context. Additionally, a fourth residual category was added not to limit coding. Next, to test the bucket themes, the two authors coded one transcript each for any references to challenges, preferences, dependencies, and configurations in relation to AI and their clinical practice. After this test, the fourth bucket theme was renamed to technical dependencies. The first author coded the remaining transcripts following the same directions. The two authors met weekly to discuss the coverage of the coding and future conceptualisation of themes. The themes were created within their respective bucket themes based on their grounding in the clinical context. Importantly, the division of codes between the bucket themes was never final and was used only to support analysis of the significant amount of codes (n=260). Through discussion, reflection on data across the interventions, and fieldwork experience, the authors iteratively clarified themes and reorganised data, moving away from the original bucket themes (while maintaining their initial assignment known). This interpretative work was conducted twice, creating ten reflective themes. The ten themes were framed as dependencies conditioning four specific design decisions that formed an AI-based support design space.

4 CONFIGURING FOUR TECHNICAL DIMENSIONS OF CLINICALLY USEFUL RADIOLOGICAL AI

We identified ten dependencies that emerge from the social dimensions of clinical AI and condition the configuration of four technical dimensions of clinical AI for radiology (see Fig. 4). Each of

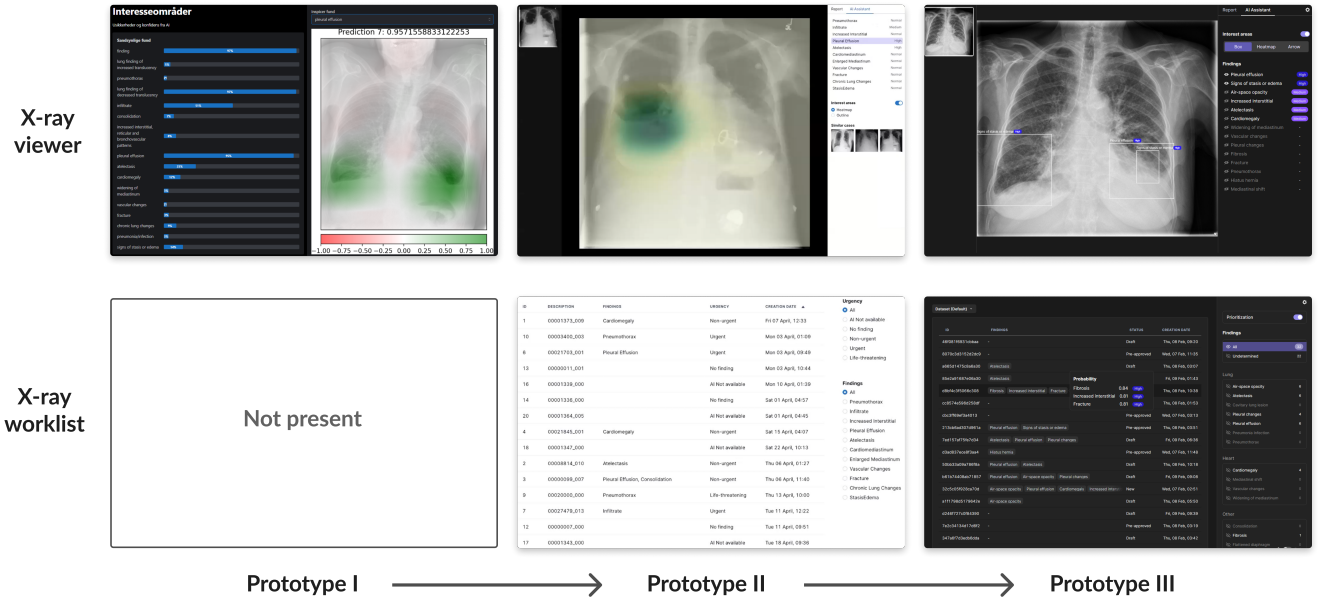

Figure 3: The collage showcases three iterations of the AI prototype, including two screens: the X-ray viewer and the X-ray worklist.

the technical dimensions needs to be configured in relation to the local clinical context to achieve clinical usefulness. In this section, we will briefly explain the social dimensions of clinical AI to then explore in-depth the conceptualised dependencies.

4.1 Social Dimensions of Clinical AI

Medical knowledge. This dimension includes concepts and definitions relevant to the medical domain addressed by the innovated AI-based system, for example, the meaning of radiological findings detected by our AI-based system. Familiarity with them supports meaningful collaboration between designers, developers, and medical professionals and reduces the risk of incorrect assumptions throughout the innovation process.

Clinic type. This social dimension addresses types of clinical sites. Imaging clinics, general hospitals, and specialised hospitals provide unique healthcare services and, thus, cater to the needs of patients with different conditions. Moreover, the type of clinical site determines the available resources, the speciality of medical professionals working there, their workflows, and their goals.

User expertise level. All medical professionals have different domain expertise. This is evident when comparing junior to senior medical professionals. However, it was also observed between board-certified radiologists. The level of expertise also determines the workload and clinical responsibilities.

Patient context. This context encompasses the current location of a patient (in or out of a hospital) and their medical history. Patients are the centre of medical work. Their health and well-being are the priority. Thus, by extension, any system supporting healthcare professionals should support patients and depend on their context.

User situation. This dimension pertains to the workload, available time, and resources of medical professionals. While the other four dependencies describe relatively stable medical practice, situational context introduces a temporal factor to the work done and may affect the priorities of medical professionals.

4.2 AI Functionality

Which AI functionality should the system provide? Answering this question defines this technical dimension. The functionalities explored during design interventions (prioritisation and decision support, see, Fig. 5) were linked to the AI model developed for the project this study was a part of. We explored the conditions for these functionalities to provide clinical value and propose a third functionality: quality assurance, which originated during the design interventions.

Dependency 1: AI functionality depends on clinic type. Each clinical site has different (1) positions within the healthcare system, (2) amounts of resources available, and (3) workloads related to the size of a clinic. This is why it is important to ensure that AI functionality is implemented in a way that makes sense for the clinical site in which it will be deployed.

First, while every radiologist puts the well-being of their patients first, the healthcare systems that they are a part of operate under different incentives. Public and private clinics face different challenges and may require adjusted AI functionalities, for example, *The number of cases in court, medical and legal cases, is way more than what you would get in the public sector. So from the medical director's office [point of view], they would want .... any small thing to be flagged so that we don't get into problems later... it would be different in K5, compared to the public sector, where even if you missed this, people are rarely taken to court but in a private setting... if they*

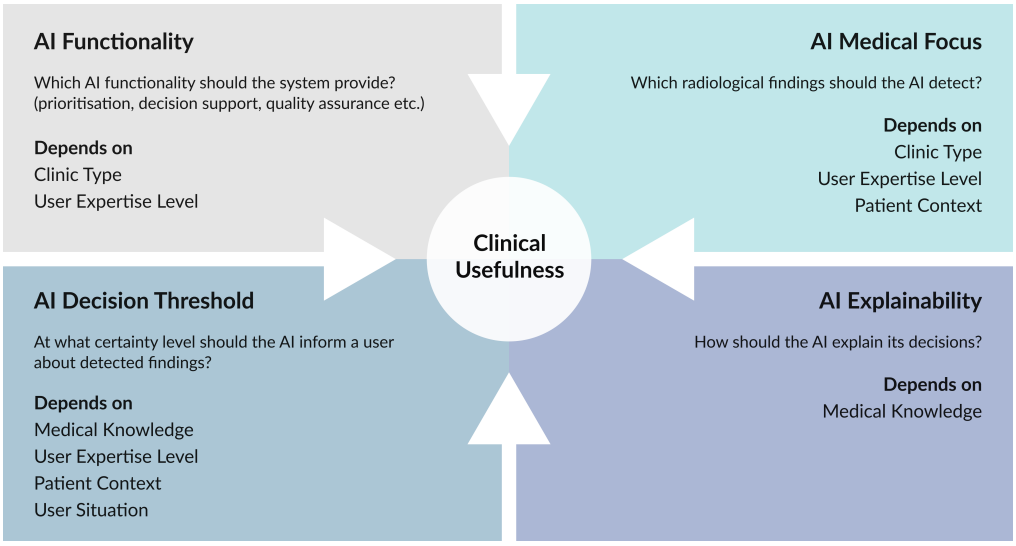

Figure 4: Configuration matrix for achieving clinical usefulness of AI. Showing how the technical AI dimensions need to be configured against the social dimensions, which they depend on.

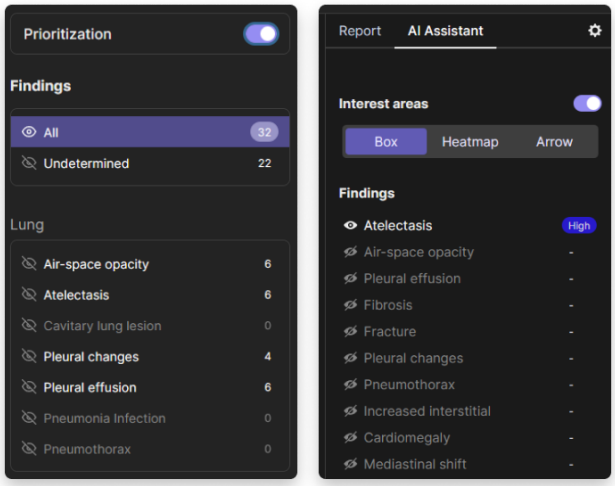

Figure 5: Two UI elements allowing the user to toggle prioritisation and decision support functionalities and to select findings on the X-ray work list and X-ray viewer.

[K5 administration] were to purchase this software, they would insist that it's set... to catch it all but you know of course this would irritate some radiologists [S18, Senior, Big general hospital, Kenya]. The difference in the prevalence of legal litigation against medical professionals in public and private healthcare centres highlighted by P05 may run along different axes in other countries. However, it is imperative for the creators of AI-based support systems to envision alternative motivations for the use of their systems and allow appropriate configuration.

Second, different clinical sites have different financial resources available. This factor, which has rarely been considered during the design of clinical AI-based systems, is a very real limitation for which functionalities will be considered worth the investment. *What is the harm in having a second opinion for each and every case? ... What is the cost? Is it a cost implication that we have to choose which images to prioritise or what?* [S18, Senior, Big general hospital, Kenya]. The different business models implemented may be detrimental to the usefulness of a system in practice. Providing decision support on all the examinations and detecting all the radiological findings may be too costly for clinics that could use such support the most, e.g., rural hospitals suffering from the lack of qualified radiologists.

Third, the clinical usefulness of AI functionalities may vary depending on the size of a clinical site, as recounted by a senior radiologist from a busy specialised hospital, *For me, the most relevant aspect of it is triage [prioritisation], but if I have five X-rays to report, then I'm not too worried because I'll get to the 5^{th} X-ray in 20 minutes. But if I have 100 X-rays to go through, I don't want to get to the 100^{th} X-ray and see that it was the one with critical findings. So in a setup where you're not very busy, I don't think it would be very useful* [S19, Senior, Specialised hospital, Kenya]. Conversely, in smaller clinics that serve mostly outpatients, implementing AI that provides quality assurance functionality would provide more value to both the radiologists and the patients. For example, *If I look at it [an examination] and [my colleague] looks at it, no one looks at it until the patient comes back four weeks later, two years later... and then "Oh, look! That's the damn thing." [e.g., a missed tumour] It could be very nice to have this second opinion* [I15, Senior, Imaging clinic, Denmark]. In this imaging clinic, radiologists, rather than being afraid of not reaching a critical patient in time, are worried about missing a critical but subtle finding, e.g., a small nodule, which may signify cancer. This means that the same AI functionality may provide useful support depending on the size of a clinic.

Dependency 2: AI functionality depends on user expertise level. The value of support in detecting findings on a medical examination decreases with increasing experience. Instead, the assigned workload increases with seniority. Thus, prioritisation and quality assurance functionalities gain importance.

Radiological AI-based decision support typically presents a list of findings detected on an examination accompanied by an XAI visualisation, as also explored in our prototypes. While this mode of support seems straightforward, it misses the reality of clinical practice. Senior radiologists spend a very short time interpreting chest X-rays. To ask them to revisit every examination to discern the validity of AI predictions is wishful. However, when discussing the potential value of AI-based decision support, they focused on quality assurance. Thus, AI should be treated not as an all-knowing peer who is going to point out every finding on an examination but as a safety net that activates only in time of need. For example, *It could read the text we write and say: "Oh, you missed that." That could be good* [I11, Senior, Imaging clinic, Denmark]. This way, the envisioned system would not require the mental effort and time to discern AI output but would inform a radiologist about potentially missed findings based on the report they were writing.

On the other hand, junior radiologists in clinical settings usually take significantly more time to report every examination. Moreover, all of their reports have to be confirmed by a senior colleague. For them, reporting serves as a primary learning exercise. In this context, they envisioned using AI support not as a quality assurance but as a new source of information used to draw their own conclusions. *I would take a look at a chest X-ray, formulate my opinion, and then see what the AI says... If it agrees... good, if it disagrees or finds something that I hadn't, I'll examine it critically... I like getting almost overwhelmed by data, and I sort it out afterwards...* [S14, Junior, Specialised hospital, Denmark]. These two perspectives highlight how workflow, workload, and the act of detecting findings on a medical examination changes with expertise. The educational value created for junior radiologists by verbose explanations of AI's predictions may become a burden for senior radiologists who expect minimal disruption to their existing workflows.

#1 Recommendation:
Enable users to select preferred AI functionality.

4.3 AI Medical Focus

Which radiological findings should the AI detect? This is where our participants, for the first time, responded, starting with "It depends..." (see Fig. 6). Let's explore how to ensure the detected findings are clinically useful in the real world.

Dependency 3: AI medical focus depends on clinic type. Different clinics take care of different types of patients suffering from different conditions. Types of patients seen in different clinical settings result in a local prevalence of observed radiological findings. As a result, a single fit-them-all system that detects an arbitrarily selected set of findings is not going to provide a similar quality of support across the different clinical contexts.

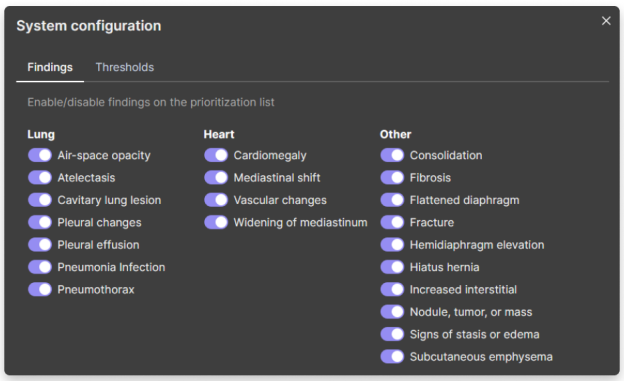

Figure 6: A configuration panel allowing users to select radiological findings detected by the AI model.

Imaging clinics and general hospitals usually examine patients referred by general practitioners. Of such patients the majority of the examinations are deemed "normal" or with findings related to infections. Hospitals with emergency departments may observe an increased prevalence of trauma-related findings, whereas specialised hospitals of post-operative, oncological, and chronic nature, as exemplified in these quotes: *That depends on the setting. If you're in a private clinic, most of the X-rays are normal... [I11] If there's something wrong, that could be pneumonia or a tumour, but usually, it's pneumonia when you go out to a private clinic* [I15, Senior, Specialised Hospital / Imaging Clinic, Denmark]. However, detecting pneumonia would not bring value in other types of settings as highlighted by P05: *If you're working in a trauma centre, the number of critical findings [e.g. pneumothorax or hemothorax] would definitely be more than in [K5], where most of the time it's just coughs and fever* [S18, Senior, Big General Hospital, Kenya]. We argue that to deliver a clinically useful AI-based system for radiologists, it is imperative to understand the local population served by the clinical site where the system is implemented. Otherwise, the developers may risk deploying a system detecting findings that may be objectively relevant to patient management yet not prevalent at the deployment site.

Dependency 4: AI medical focus depends on user expertise level. Junior radiologists may interpret a single X-ray for up to tens of minutes. Whereas, according to our senior participants, interpreting a chest X-ray takes around 1 to 2 minutes. This means that with experience, many findings become "obvious" and are no feat to detect. When discussing the decision support functionality of our prototypes and previous systems that our participants had piloted, the common complaint related to the detection of "obvious" radiological findings, which took additional time to discern.

If it's an obvious finding, we'll see that one quickly, and we all agree on it. The problem comes when it's something more subtle [I06, Senior, Big General Hospital, Kenya]. Detecting the difficult or "subtle" radiological findings is where the value lies for senior radiologists. However, the less experienced, the more support a radiologist may accept. This was captured by P01: *Maybe it'll help the resident radiologist in the first or second year, but I don't think*

it will help a specialised radiologist with experience because once we can have a look, we can't miss something like this [I01, Senior, Small General Hospital, Kenya]. This means that in order to support different radiologists in practice, AI-based systems may need to allow users to select findings to receive support with. Without such configuration, discerning AI predictions regarding "obvious" findings, even when true, would result in more time spent and annoyance.

Dependency 5: AI medical focus depends on patient context. Radiologists are not interpreting medical imaging to find every possible finding. Rather, they are interpreting them to help the ordering clinicians take action in patient management. Such actions usually occur when a new condition is being diagnosed, or a patient's health may be at risk. However, the clinical meaning of certain radiological findings depends on the location of a patient. This means that a finding may be expected when observed in an examination of a patient who is admitted to a hospital. Whereas the same finding observed in an examination of a patient who is not admitted to a hospital may warrant immediate action.

Our participants stressed that useful prioritisation should consider patients' medical history to filter out already-known findings, which our prototype could not do. *But how urgent is it? We know that pneumothorax has decreased. It's a big heart, but it's much smaller than it was a week ago. It has [pleural] effusion, but much, much less than it was a week ago. That's the thing we miss with this* [I11, Senior, Specialised Hospital / Imaging Clinic, Denmark]. In this quote, P11 explains that the examination they looked at may not be urgent at all despite the fact that the AI correctly detected three findings, one of them (pneumothorax) being life-threatening. These findings would not be urgent if they were already known to the ordering clinician. In such a case, the patient would have already been undergoing treatment, and this examination's sole purpose was to control its progress. In specialised hospitals and bigger general hospitals, patients often have taken several X-rays to monitor the progress of treatment. This means that the same findings, but of different severity, will be visible on their examinations. The ability to assess the detected findings in the light of patient history is crucial to correctly prioritise findings that warrant clinical action.

When looking at radiologists' work from the perspective of contributing to the broader clinical work, it is counterproductive to prioritise findings that clinicians taking care of a patient are already aware of. In other words, a radiological finding may be relevant to detect on examinations from patients who are not admitted to a hospital, but not so much for patients currently admitted. A senior radiologist explained, *It depends on the findings, and it depends on the patient... some findings in the out-patients would be more important to be prioritised than if they're in-house. Because if they're in-house, then I would suspect that someone not from the radiology department would have looked at them. If it's out-patient, then nobody has looked at them...* [I10, Senior, Specialised Hospital, Denmark]. Whereas, as explained by P10, patients referred from outside of a hospital are more likely to have conditions that their doctors are unaware of. Thus, the location of the patient is crucial to selecting which findings are relevant to receiving support from an AI-based system.

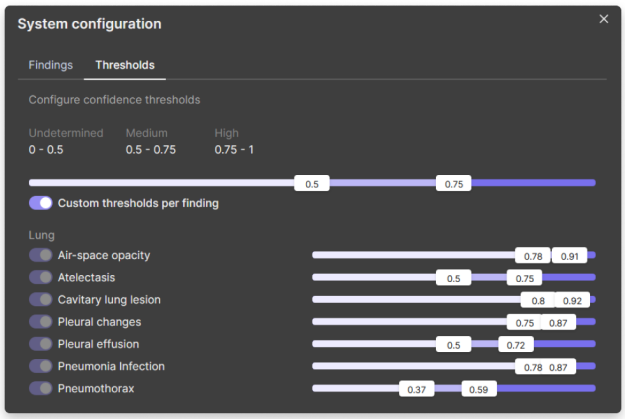

Figure 7: A configuration panel allowing users to select AI decision threshold globally and per finding. In this version, we introduced two levels for quick access depending on the user situation: high confidence and medium confidence.

#2 Recommendation:
Enable users to select radiological findings.

4.4 AI Decision Threshold

At what certainty level should the AI inform a user about detected findings? Specifying when a radiological AI-based system should inform a user about a finding is usually done by specifying a decision threshold (see Fig. 7). Selecting a specific threshold value determines the measured performance of an AI model captured by evaluation metrics like specificity, sensitivity, or positive and negative predictive values. Arguably, in practice, a decontextualised performance value is less important than the practical consequences of selecting a specific threshold level. Every time an AI model detects a finding (based on a selected threshold), a radiologist may have to take action to assess it. The balance between clinical value and additional burden is thus closely tied to how well the threshold is configured to match the local clinical context. We conceptualised four dependencies that influence the configuration of the AI decision threshold.

Dependency 6: AI decision threshold depends on medical knowledge. While some of the radiological findings are well understood across the contexts, some definitions are more subjective and their meanings change across countries. Infiltration or consolidation are two examples of radiological findings which have been found to be used differently in clinical practice in Denmark and Kenya. Moreover, some of the findings were too vague for the radiologists to decide how to assess them, for example, *I think vascular changes can mean one of two things if it's the big vessels - I think it's important to have it, e.g., if the computer can say the aorta is big... It could also be about... the small vessels, and then it's more like stasis. Then it's quite different* [I17, Junior, Specialised hospital, Denmark]. The underlying definition of a finding, in this described case, affected how P13 understood the condition and what threshold level they deemed

appropriate. Within radiology, chest X-rays are a particularly subjective modality. Due to their visual complexity, radiologists rely on their expertise to interpret the observed findings. Precise definitions used to label data the AI model was trained on are crucial to assess the predictions.

Dependency 7: AI decision threshold depends on user expertise level. Often, junior and senior radiologists are juxtaposed as two groups of AI support end-users with different needs. This is also visible in the strategy for selecting thresholds.

When used by junior radiologists, both junior and senior radiologists (who supervised them) leaned towards accepting AI predictions only with a high degree of certainty. As explained before, the interpretation of chest X-rays is uniquely subjective. It takes experience to report them with a high degree of certainty. In this context, an uncertain AI prediction would jeopardise the learning process and introduce more confusion, resulting in more work for the junior students and their supervisors.

On the other hand, allowing senior radiologists to set the threshold for different findings according to their personal preferences could entice them to utilise the system in their own way. P05, who also had senior administrative experience, explained that senior radiologists do not always have the same level of expertise and may need different levels of support. *This would be amazing. I wouldn't want to do it [adjust threshold] at the administrative level because ... not all the radiologists in the department have the same capabilities. So I'd rather let people set it for themselves* [S18, Senior, Big general hospital, Kenya]. By enabling users to select the AI decision threshold on their own, they could build trust by incrementally including AI in their own practice.

Dependency 8: AI decision threshold depends on patient context. Diverging from a fixed threshold level defined at a system level towards finding-level threshold specification may boost the clinical usefulness of AI-based systems for radiology. A finding-level threshold specification would allow radiologists to stratify which findings in a given context are more relevant.

They could do it by lowering the threshold. A lower threshold would be associated with a higher rate of false positive prediction for that particular radiological finding. Thus, more work for radiologists. However, for a subset of findings, our participants were willing to accept more false positive detections if it would benefit their patients. *For pneumothorax, I would probably lower the threshold because you would want to find every pneumothorax there is. But for some other stuff, like fibrosis, I would probably have a higher threshold because that's not critical* [S13, Senior, Specialised hospital, Denmark]. Based on design interventions with the third prototype, our participants saw a utility in such fine-grained configuration. *I think the relevance of certainty [threshold] is the clinical implication of the diagnosis. So, something like a pneumothorax needs some form of intervention... whereas on a suspected infection, a clinician may go ahead and treat it even if the X-ray is normal. So that's why it may not be such a big deal whether I call a pneumonia or not. Whereas a pneumothorax might need a chest tube insertion. It's a do-or-die call* [S19, Senior, Specialised hospital, Kenya]. As shown in this quote, the clinical implications for a patient made radiologists more accepting of false positives. Meanwhile, findings that were less severe or that could be discerned using other indicators, e.g., clinical indicators (cough, fever) to decide on pneumonia diagnosis, were less preferred to lower the threshold. This suggests that configuring AI decision threshold on a finding level could reduce the workload associated with false positive predictions and help focus AI support.

Dependency 9: AI decision threshold depends on user situation. Radiologists' approach to AI support changes with time. In this paper, we uncovered two temporal aspects that affected how radiologists thought of configuring AI decision thresholds: the time spent using the system and the rhythm of clinical work.

One of the common comments when discussing the threshold with our participants was about its arbitrary nature. Radiologists wondered what the real-life consequences of changing the threshold would be. Based on these concerns, our final prototype included an estimate of false positive predictions. These values, while more relatable, were still considered difficult to imagine in real practice both for senior and junior radiologists. *I mean, it's a bit arbitrary at this moment because you don't have any idea what the effect is* [I17, Junior, Specialised hospital, Denmark]. *It would be nice to be able to adjust this... try all this out and see in real life how many cases it's missing or over-calling* [S19, Senior, Specialised hospital, Kenya]. These quotes highlight that such essential development tasks as selecting a threshold have little to no basis in clinical practice. They uncover a need for a better translation between the domains of AI and Health to support meaningful configuration. Currently, this translation has to be conducted through real-world experimentation in the final context of use. This way, medical professionals may gain a practical understanding of what the changes to the threshold mean and further purposefully and consciously adjust it to fit their work.

The second temporal aspect of selecting an appropriate threshold relates to the routine of end users. Radiologists saw an advantage in adjusting the threshold depending on their workload. For example, a specialised radiologist from a busy specialised hospital mentioned, on *Fridays, we tend to be more active because if you leave a long list on Friday, the turnaround time will be way longer - there is very low coverage over the weekend [few on-call doctors]... and then Monday tends to be very busy* [I04, Senior, Specialised hospital, Kenya]. During this conversation, the radiologist concluded that lowering the threshold could help them ensure that no examinations with critical findings were left to be reported after the weekend. These two aspects highlight that what radiologists consider a useful level of detection (including false positive predictions) may vary throughout the use.

#3 Recommendation:
Enable users to adjust AI thresholds.

4.5 AI Explainability

How should the AI explain its decisions? Understanding AI predictions supports building trust towards AI-based systems. In this study, we explored three visual ways of explaining AI predictions: heat maps, bounding boxes, and arrows (see Fig. 8). We discovered that no single method can support the explainability of all the radiological findings.

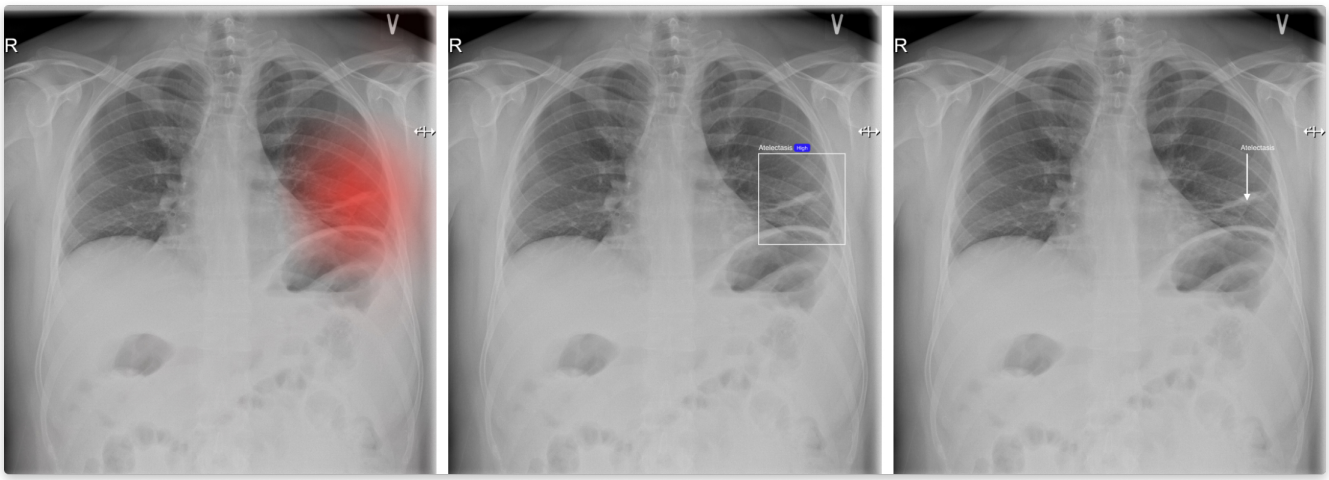

Figure 8: Different XAI methods available in the prototype. From left to right: gradient overlay, bounding box, arrow.

Dependency 10: AI explainability depends on medical knowledge. The visual appearance of radiological findings dictates the best way to highlight them for radiologists for inspection. Radiologists discern between different radiological findings based on their visual appearance. Their presentation ranges from barely visible nodules to diffused opacities (areas of less transparency) present across both lungs. The breadth of visual impressions suggests the need for flexibility, *I think that both ways of displaying the findings are fine, but for different pathologies. I mean, the heat map makes sense in this case for pneumothorax because it's a very extensive finding. And for the fracture, it makes sense to see it with a box, whereas the heat map doesn't make that much sense. It becomes too blurry...* [S13, Senior, Specialised hospital, Denmark]. Radiologists preferred bounding boxes for more contained findings, whereas the more diffused, the more inclined they were towards the heat map. An important factor when designing XAI for chest X-rays is allowing for inspection of the underlying examination. The main purpose of XAI is to direct radiologists' attention to the detected findings. To assess the validity of a prediction, radiologists have to inspect the examination itself without additional overlays.

#4 Recommendation:
Enable users to choose the most suitable XAI method.

5 DISCUSSION

In this paper, we investigated how to design AI for clinical usefulness in different clinical contexts of radiology practice. Based on extended research through design study, we provided four practical recommendations on addressing ten dependencies emerging from the social dimensions of clinical practice (Fig. 9). By engaging radiologists in two different countries from the Global North and Global South, respectively, we found that the radiologists' practices in Kenya and Denmark were rather similar, possibly due to resemblances in medical education, scientific models, and ethics. The social dimensions derived from this study therefore orient towards similarities that cut across country specifics. However, the types of healthcare systems and present IT infrastructures differed quite substantially and need to be accounted for during AI innovation, which goes beyond this study. In this section, we will discuss how these recommendations may be enacted during the innovation process of clinical AI for different clinical contexts.

5.1 Enable users to select preferred AI functionality

Configuring clinical AI-based support systems to suit local environments is essential, as one-size-fits-all approaches often fail to address their unique needs [34, 39, 58, 89]. In this study, we discovered how social dimensions of clinical practice condition what kind of AI functionality is considered useful. We argue that for AI to match local requirements, it needs to be configured throughout the innovation process with the intended context of use in mind, i.e., the expertise of the end-users and the work performed in their clinical site. This is especially relevant, as clinical AI is often afflicted by the problem of late realisation [33, 88].

When addressing expertise-related needs for support, previous research in radiology showed that AI-based systems have different effects on junior and senior radiologists [95]. Even more, Tong et al. [78] investigated two strategies, what they called "optimised" and "all-AI", for AI support of junior and senior radiologists in thyroid nodule management. They reported that the best results were obtained when the type of support was configured to the expertise of the radiologist. However, our study showed that personal preferences play a deciding factor only if the AI functionality is appropriate in the context of the local clinic. Selecting the best way to prioritise findings will not make sense if there are only a few examinations to prioritise to begin with. AI-based systems should be designed to respond to fit the utility gap in a clinic and then be configured to the varying needs and preferences of different end-users, depending on their level of experience, knowledge, and confidence.

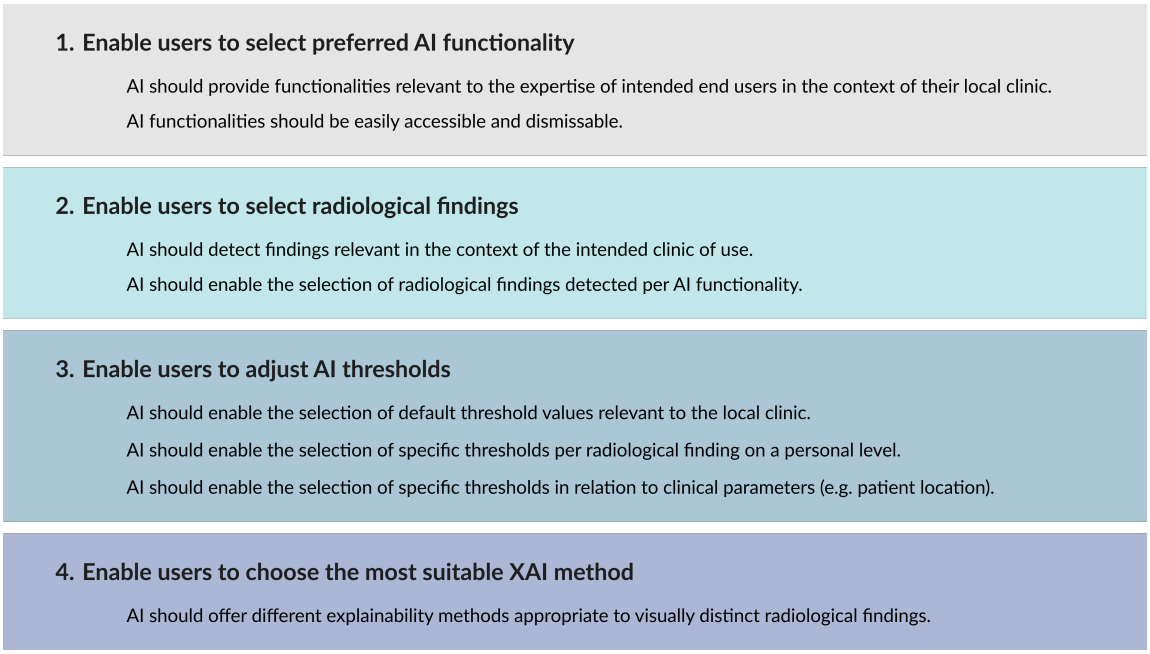

Figure 9: Four design recommendations on how to achieve clinically useful AI-based systems. Accompanied by more in-depth considerations.

The personal configuration of functionality also captures the integration - a famously difficult task when innovating clinical AI [58, 77, 90]. Many AI-based systems fail in practice due to providing support at the wrong time [9, 19, 32, 37]. Some of the AI integrations introduce a new step in the practice. A step that sometimes cannot be skipped [9]. This study shows, seconding previous research, that the integration of AI into work practices has to be flexible [16]. Clinical work is always changing, and so are the needs for AI support. Thus, we recommend that clinical end-users should be in control of which AI functionalities are a part of their current routine.

5.2 Enable users to select radiological findings

The innovation of AI-based systems is often initiated and defined by technical opportunities, e.g., access to medical data [70, 71, 87, 94]. As such, the medical and social aspects of the systems are sometimes addressed only after the technology has gone through several rounds of development [2]. This inadvertently means that certain assumptions about the medical focus are made [92]. Present radiological AI models tend to detect findings relevant to the local radiologists involved in the data creation process [38, 54, 83]. However, we showed that the prevalence and clinical meaning of radiological findings varies based on the clinic type and patient context. This affects the usefulness of clinical systems in different settings and their transferability [55, 92]. Thus, it is critical to investigate the intended clinical context of use prior to deciding on the medical focus of the AI-based system and to allow medical professionals to set the scope of support relevant to them and their practice.

Moreover, the clinical meaning of radiological findings is tied to the patient context and not only the type of medical condition, i.e., a radiological finding expected in an in-patient examination can be life-threatening when found in an out-patient one. This discovery deepens our understanding of how medical professionals make decisions and in what situations they may need AI support. This finding contrast with systems where certain radiological findings are consistently considered urgent in patient care [86]. We suggest that linking clinical information about a patient with detected findings may better reflect radiologists' actual decision-making practices and result in improved usefulness of the AI-based system. This is why we recommend including clinical information in conjunction with AI predictions to better respond to the real-world needs of medical professionals.

5.3 Enable users to adjust AI thresholds

Selecting AI decision threshold has significant ethical [12], performance [65], and clinical [81] consequences for AI-based systems, and it has been a notable research topic in the AI and Health communities. Recently, it gained footing in the HCI design community. Kocielnik et al. [44] explored how the decision threshold affects the number of false positive and false negative predictions, significantly altering users system perception. While from the technical point of view, the accuracy may be the same, the distribution of false positives and false negatives may have severe clinical consequences. Our participants warned that false positive predictions require additional time and resources to discern and that the potential benefits of AI often do not justify this additional cost, resulting in the failure of the AI-based systems in clinical practice [5, 7, 50, 76].

However, until AI reaches 100% accuracy, false positive predictions are the reality of AI-based systems. Improving performance is only one way of addressing them. In this paper, we offer another

outlook, namely, addressing the cost-benefit ratio of AI predictions. This ratio is not static. Just like clinical practice, it fluctuates and depends on time, workload, known critical cases, and available resources. In certain situations, medical professionals may accept more false positive predictions, e.g. when making sure that there are no critical findings in a queue of examinations that will not be looked at over the weekend. This means that regardless of how well an AI decision threshold is preset, AI will not provide the same value throughout its use in clinical practice. Supporting end-users in configuring the AI decision threshold depending on their local needs can improve the clinical usefulness of AI-based systems. Thus, designers and developers should enable end-user configuration of decision thresholds in clinician-facing AI systems.

5.4 Enable users to choose the most suitable XAI method

It has been long established that explainable AI fosters trust and increases the usefulness of the predictions [24, 39]. Especially in the healthcare domain, the reasoning and explanations are sometimes more valuable to end users than the predictions themselves [19, 50] or can lead to envisioning new ways of using an AI-based system altogether [40]. However, simply revealing the decision-making process of machines to humans is not enough to provide useful explanations [52]. Instead, our study suggests that for XAI methods to be effective in explaining medical conditions, they must be configured to how medical professionals assess those conditions. This means that even proven methods used in medical imaging, like heat maps or bonding boxes, when used to highlight incompatible conditions, may cause confusion and require additional work to discern. To this end, we recommend that to ensure the clinical usefulness of XAI methods, they should be configurable in accordance with medical knowledge.

6 LIMITATIONS AND FUTURE WORK

This work is not without its limitations. As explored in this paper, when interacting with the prototype, radiologists envisioned support functionalities like quality assurance through the assessment of written reports against AI's interpretation of findings on a chest X-ray. This functionality was outside of the prototyped prioritisation and decision support. This choice was dictated by the capabilities of the underlying AI model and the innovation direction of the greater project this study was a part of. We believe that this mismatch perfectly exemplifies the difficulty of innovating clinically useful AI-based systems and motivates further research into a meaningful configuration of AI-based systems, especially at the defining early stages of work. In addition, the study did not take into consideration the differences in clinicians' attitudes and expectations towards AI as well as the collaborative aspects of reporting, which may be worth considering in studies that follow up on designing for clinical usefulness.

We also acknowledge the limited variability of clinical sites in Denmark compared to the visited sites in Kenya due to difficulties gaining access. While it is a strength of the study that medical professionals from very different countries participated, future studies need to further explore how geographic and cultural differences play out in regard to successfully designing for and transferring AI-based systems to entirely different healthcare systems and IT infrastructures. Moreover, this project commenced before large language models experienced a performance leap. We believe that their ability to parse and produce text may be an opportune, although challenging avenue for AI support to explore [91].

7 CONCLUSIONS

Innovating clinical AI-based systems is a challenging task. By investigating design interventions conducted with radiologists across diverse clinical contexts in Denmark and Kenya, we identified four key technical dimensions that require careful configuration: AI functionality, AI medical focus, AI decision threshold, and AI explainability. To support the innovation of clinically useful AI-based systems, we derived four concrete recommendations of what we propose to call "configurable AI" pertaining to the four key technical dimensions.

Moreover, we explored how dependencies originating from the social dimensions of local clinical practice condition the clinical usefulness of the uncovered technical dimensions. AI functionalities (e.g., prioritisation or decision support) should be configured to provide value in the intended type of clinical site and to match the level of medical expertise of end users. AI medical focus (the detected findings in radiology-focused systems) should be configured in relation to the patient's context, the level of medical expertise of the end-users, and the type of clinical site. The AI decision threshold should be configured according to the medical knowledge (e.g., the clinical meaning of radiological findings), the patient's context, the level of medical expertise of the end users, and the user situation (e.g. time of day). Finally, the explainable AI should be configurable in accordance with medical knowledge to provide maximum value to the end-users.

Our findings highlight the need for designers and developers to consider these dependencies throughout the innovation process, both before-use and in-use, to ensure that AI-based systems are effectively configured to meet the needs and requirements of their intended clinical contexts. By adhering to these recommendations and considering the dependencies uncovered in our study, designers and developers can contribute to the successful innovation of clinically useful AI-based systems in radiology, ultimately improving patient care and clinical outcomes.

REFERENCES

[1] Scott J. Adams, Robert D.E. Henderson, Xin Yi, and Paul Babyn. 2021. Artificial Intelligence Solutions for Analysis of X-ray Images. *Canadian Association of Radiologists Journal* 72, 1 (2 2021), 60–72. https://doi.org/10.1177/0846537120941671/ASSET/IMAGES/LARGE/10.1177{_}0846537120941671-FIG2.JPEG

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019*. 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

[3] Tariq Andersen, Joachim Halse, and Jonas Moll. 2011. Design interventions as multiple becomings of healthcare. In *Proceedings of the 4th Nordic Design Research Conference (Nordes'11), Helsinki, Finland, May 29–31, 2011, Helsinki: Aalto University*. 11–20.

[4] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* 25, 6 (6 2019), 954–961. https://doi.org/10.1038/s41591-019-0447-x

[5] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. 2023. "If I Had All the Time in the World": Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3544548.3581513

[6] Ellen Balka and Ina Wagner. 2006. Making things work: dimensions of configurability as appropriation work. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (Banff, Alberta, Canada) *(CSCW '06)*. Association for Computing Machinery, New York, NY, USA, 229–238. https://doi.org/10.1145/1180875.1180912

[7] Sally L. Baxter, Jeremy S. Bass, and Amy M. Sitapati. 2020. Barriers to Implementing an Artificial Intelligence Model for Unplanned Readmissions. *ACI Open* 04, 02 (7 2020), e108–e113. https://doi.org/10.1055/s-0040-1716748

[8] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. 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 (CHI '20)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[9] Natalie C Benda, Lala Tanmoy Das, Erika L Abramson, Katherine Blackburn, Amy Thoman, Rainu Kaushal, Yongkang Zhang, and Jessica S Ancker. 2020. "how did you get to this number?" Stakeholder needs for implementing predictive analytics: A pre-implementation qualitative study. *Journal of the American Medical Informatics Association* 27, 5 (5 2020), 709–716. https://doi.org/10.1093/jamia/ocaa021

[10] Arngeir Berge, Frode Guribye, Siri-Linn Schmidt Fotland, Gro Fonnes, Ingrid H. Johansen, and Christoph Trattner. 2023. Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, Vol. 23. ACM, New York, NY, USA, 1339–1352. https://doi.org/10.1145/3563657.3596110

[11] Shlomo Berkovsky and Enrico Coiera. 2023. Moving beyond algorithmic accuracy to improving user interaction with clinical AI. *PLOS Digital Health* 2, 3 (3 2023), e0000222. https://doi.org/10.1371/journal.pdig.0000222

[12] Jonathan Birch, Kathleen A. Creel, Abhinav K. Jha, and Anya Plutynski. 2022. Clinical decisions using AI must consider patient values. *Nature Medicine* 28, 2 (2 2022), 229–232. https://doi.org/10.1038/s41591-021-01624-y

[13] Ann Blandford, Jo Gibbs, Nikki Newhouse, Olga Perski, Aneesha Singh, and Elizabeth Murray. 2018. Seven lessons for interdisciplinary research on interactive digital health interventions. *DIGITAL HEALTH* 4 (2018), 205520761877032. https://doi.org/10.1177/2055207618770325

[14] Jeanette Blomberg and Helena Karasti. 2013. Reflections on 25 years of ethnography in CSCW. *Computer supported cooperative work (CSCW)* 22 (2013), 373–423.

[15] Jeanette Blomberg, Lucy Suchman, and Randall H. Trigg. 1996. Reflections on a work-oriented design project. *Hum.-Comput. Interact.* 11, 3 (sep 1996), 237–265. https://doi.org/10.1207/s15327051hci1103_3

[16] Claus Bossen and Kathleen H. Pine. 2023. Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists. *ACM Transactions on Computer-Human Interaction* 30, 2 (4 2023), 1–29. https://doi.org/10.1145/3569892

[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[18] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Vol. 19. ACM, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581251

[19] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[20] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–24. https://doi.org/10.1145/3359206

[21] Michelle Campbell, R Fitzpatrick, A Haines, A L Kinmonth, P Sandercock, D Spiegelhalter, and P Tyrer. 2000. Framework for design and evaluation of complex interventions to improve health. *BMJ* 321, 7262 (9 2000), 694–6. https://doi.org/10.1136/bmj.321.7262.694

[22] Enrico Coiera. 2019. The last mile: Where artificial intelligence meets reality. , 16323 pages. https://doi.org/10.2196/16323

[23] Kathrin M Cresswell, Ann Blandford, and Aziz Sheikh. 2017. Drawing on human factors engineering to evaluate the effectiveness of health information technology. , 309–315 pages. https://doi.org/10.1177/0141076817712252

[24] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*, Vol. 12. ACM, New York, NY, USA, 1591–1602. https://doi.org/10.1145/3461778.3462131

[25] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 1639–1656. https://doi.org/10.1145/3531146.3533221

[26] Paul Dourish. 1997. Accounting for system behavior: representation, reflection, and resourceful action. In *Computers and design in context*. 145–170.

[27] Paul Dourish. 2003. The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents. *Computer Supported Cooperative Work (CSCW)* 12, 4 (12 2003), 465–490. https://doi.org/10.1023/A:1026149119426

[28] Paul N. Edwards, Steven J. Jackson, Geoffrey C. Bowker, and Cory P. Knobel. 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. *Self-published*. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/Understand?sequence=3

[29] Madeleine Clare Elish and Elizabeth Anne Watkins. 2020. *Repairing Innovation: A Study of Integrating AI in Clinical Care*. Technical Report. Data & Society. https://datasociety.net/wp-content/uploads/2020/09/Repairing-Innovation-DataSociety-20200930-1.pdf

[30] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, Vol. 22. ACM, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[31] Aimilia Gastounioti, Vasileios Kolias, Spyretta Golemati, Nikolaos N. Tsiaparas, Aikaterini Matsakou, John S. Stoitsis, Nikolaos P.E. Kadoglou, Christos Gkekas, John D. Kakisis, Christos D. Liapis, Petros Karakitsos, Ioannis Sarafis, Pantelis Angelidis, and Konstantina S. Nikita. 2014. CAROTID - A web-based platform for optimal personalized management of atherosclerotic patients. *Computer Methods and Programs in Biomedicine* 114, 2 (2014), 183–193. https://doi.org/10.1016/j.cmpb.2014.02.006

[32] Jennifer C. Ginestra, Heather M. Giannini, William D. Schweickert, Laurie Meadows, Michael J. Lynch, Kimberly Pavan, Corey J. Chivers, Michael Draugelis, Patrick J. Donnelly, Barry D. Fuchs, and Craig A. Umscheid. 2019. Clinician Perception of a Machine Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock. *Critical care medicine* 47, 11 (11 2019), 1477–1484. https://doi.org/10.1097/CCM.0000000000003803

[33] Fabien Girardin and Neal Lathia. 2017. When user experience designers partner with data scientists. In *AAAI Spring Symposium - Technical Report*, Vol. SS-17-01 -. 376–381. www.aaai.org

[34] Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang 'Anthony' Chen. 2021. Lessons Learned from Designing an AI-Enabled Diagnosis Tool for Pathologists. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (4 2021), 1–25. https://doi.org/10.1145/3449084

[35] Joachim Halse and Laura Boffi. 2016. Design interventions as a form of inquiry. In *Design Anthropological Futures*, Rachel Charlotte Smith, Kasper Tang Vangkilde, Mette Gislev Kjærsgaard, Ton Otto, Joachim Halse, and Thomas Binder (Eds.). Bloomsbury, London. https://adk.elsevierpure.com/en/publications/design-anthropological-futures

[36] Morten Hertzum and Jesper Simonsen. 2019. Configuring information systems and work practices for each other: what competences are needed locally? *International Journal of Human-Computer Studies* 122 (2019), 242–255.

[37] Judd E. Hollander, Keara L. Sease, Dina M. Sparano, Frank D. Sites, Frances S. Shofer, and William G. Baxt. 2004. Effects of neural network feedback to physicians on admit/discharge decision for emergency department patients with chest pain. *Annals of Emergency Medicine* 44, 3 (9 2004), 199–205. https://doi.org/10.1016/j.annemergmed.2004.02.037

[38] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (1 2019), 590–597. http://arxiv.org/abs/1901.07031

[39] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445385

[40] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Trans.*

Comput. Healthcare 1, 1 (3 2020). https://doi.org/10.1145/3344258

[41] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3517533

[42] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* 17, 1 (2019), 1–9. https://doi.org/10.1186/s12916-019-1426-2

[43] Finn Kensing and Andreas Munk-Madsen. 1993. PD: structure in the toolbox. *Commun. ACM* 36, 6 (6 1993), 78–85. https://doi.org/10.1145/153571.163278

[44] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[45] Morten Kyng and Lars Mathiassen. 1997. *Computers and design in context*. MIT Press. https://cir.nii.ac.jp/crid/1130282270228481664

[46] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (9 2023), 1–22. https://doi.org/10.1145/3610218

[47] Constance D. Lehman, Robert D. Wellman, Diana S. M. Buist, Karla Kerlikowske, Anna N. A. Tosteson, and Diana L. Miglioretti. 2015. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Internal Medicine* 175, 11 (11 2015), 1828. https://doi.org/10.1001/jamainternmed.2015.5231

[48] Dana Li, Lea Marie Pehrson, Carsten Ammitzbøl Lauridsen, Lea Tøttrup, Marco Fraccaro, Desmond Elliott, Hubert Dariusz Zając, Sune Darkner, Jonathan Frederik Carlsen, and Michael Bachmann Nielsen. 2021. The Added Effect of Artificial Intelligence on Physicians' Performance in Detecting Thoracic Pathologies on CT and Chest X-ray: A Systematic Review. *Diagnostics* 11, 12 (11 2021), 2206. https://doi.org/10.3390/diagnostics11122206

[49] Xiaoxuan Liu, Samantha Cruz Rivera, Livia Faes, Lavinia Ferrante Di Ruffano, Christopher Yau, Pearse A Keane, Hutan Ashrafian, Ara Darzi, Sebastian J Vollmer, Jonathan Deeks, Lucas Bachmann, Christopher Holmes, An Wen Chan, David Moher, Melanie J. Calvert, and Alastair K. Denniston. 2019. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nature Medicine* 25, 10 (10 2019), 1467–1468. https://doi.org/10.1038/s41591-019-0603-3

[50] Stina Matthiesen, Søren Zöga Diederichsen, Mikkel Klitzing Hartmann Hansen, Christina Villumsen, Mats Christian Højbjerg Lassen, Peter Karl Jacobsen, Niels Risum, Bo Gregers Winkel, Berit T Philbert, Jesper Hastrup Svendsen, and Tariq Osman Andersen. 2021. Clinician Preimplementation Perspectives of a Decision-Support Tool for the Prediction of Cardiac Arrhythmia Based on Machine Learning: Near-Live Feasibility and Qualitative Study. *JMIR Human Factors* 8, 4 (11 2021), e26964. https://doi.org/10.2196/26964

[51] Mohammad H Rezazade Mehrizi, Simon H Gerritsen, Wouter M de Klerk, Chantal Houtschild, Silke M H Dinnessen, Luna Zhao, Rik van Sommeren, and Abby Zerfu. 2023. How do providers of artificial intelligence (AI) solutions propose and legitimize the values of their solutions for supporting diagnostic radiology workflow? A technography study in 2021. *European radiology* 33, 2 (2 2023), 915–924. https://doi.org/10.1007/s00330-022-09090-x

[52] Cecily Morrison, Kit Huckvale, Bob Corish, Richard Banks, Martin Grayson, Jonas Dorn, Abigail Sellen, and Sân Lindley. 2018. Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (6 2018), 1–28. https://doi.org/10.1145/3181670

[53] Ju Gang Nam, Sunggyun Park, Eui Jin Hwang, Jong Hyuk Lee, Kwang-Nam Jin, Kun Young Lim, Thienkai Huy Vu, Jae Ho Sohn, Sangheum Hwang, Jin Mo Goo, and Chang Min Park. 2019. Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology* 290, 1 (1 2019), 218–228. https://doi.org/10.1148/radiol.2018180237

[54] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q Tran, Dung B. Nguyen, Dung D Le, Chi M Pham, Hang T T Tong, Diep H Dinh, Cuong D Do, Luu T Doan, Cuong N. Nguyen, Binh T Nguyen, Que V. Nguyen, Au D Hoang, Hien N Phan, Anh T Nguyen, Phuong H Ho, Dat T Ngo, Nghia T Nguyen, Nhan T Nguyen, Minh Dao, and Van Vu. 2020. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. (2020). https://vindr.ai/vinlabhttp://arxiv.org/abs/2012.15029

[55] John Nolan, Peter McNair, and Jytte Brender. 1991. Factors influencing the transferability of medical decision support systems. *International Journal of Bio-Medical Computing* 27, 1 (1 1991), 7–26. https://doi.org/10.1016/0020-7101(91)90018-A

[56] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in Healthcare: Challenges Appearing in the Wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–5. https://doi.org/10.1145/3411763.3441347

[57] Cécile Petitgand, Aude Motulsky, Jean Louis Denis, and Catherine Régis. 2020. Investigating the barriers to physician adoption of an artificial intelligence-based decision support system in emergency care: An interpretative qualitative study. In *Studies in Health Technology and Informatics*, Vol. 270. IOS Press, 1001–1005. https://doi.org/10.3233/SHTI200312

[58] Kathleen H. Pine and Yunan Chen. 2020. Right Information, Right Time, Right Place: Physical Alignment and Misalignment in Healthcare Practice. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376818

[59] Rob Procter, Peter Tolmie, and Mark Rouncefield. 2023. Holding AI to Account: Challenges for the Delivery of Trustworthy AI in Healthcare. *ACM Transactions on Computer-Human Interaction* 30, 2 (4 2023), 1–34. https://doi.org/10.1145/3577009

[60] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine* 15, 11 (11 2018), e1002686. https://doi.org/10.1371/journal.pmed.1002686

[61] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, Alessandro Ruggiero, Anna Korhonen, Emily Jefferson, Emmanuel Ako, Georg Langs, Ghassem Gozaliasl, Guang Yang, Helmut Prosch, Jacobus Preller, Jan Stanczuk, Jing Tang, Johannes Hofmanninger, Judith Babar, Lorena Escudero Sánchez, Muhunthan Thillai, Paula Martin Gonzalez, Philip Teare, Xiaoxiang Zhu, Mishal Patel, Conor Cafolla, Hojjat Azadbakht, Joseph Jacob, Josh Lowe, Kang Zhang, Kyle Bradley, Marcel Wassin, Markus Holzer, Kangyu Ji, Maria Delgado Ortet, Tao Ai, Nicholas Walton, Pietro Lio, Samuel Stranks, Tolou Shadbahr, Weizhe Lin, Yunfei Zha, Zhangming Niu, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3, 3 (3 2021), 199–217. https://doi.org/10.1038/s42256-021-00307-0

[62] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445748

[63] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, Matthew G Wallis, Ingvar Andersson, Sophia Zackrisson, Ritse M Mann, and Ioannis Sechopoulos. 2019. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute* 111, 9 (9 2019), 916–922. https://doi.org/10.1093/jnci/djy222

[64] Santiago Romero-Brufau, Kirk D. Wyatt, Patricia Boyum, Mindy Mickelson, Matthew Moore, and Cheristi Cognetta-Rieke. 2020. A lesson in implementation: A pre-post study of providers' experience with artificial intelligence-based clinical decision support. *International Journal of Medical Informatics* 137, November 2019 (2020), 104072. https://doi.org/10.1016/j.ijmedinf.2019.104072

[65] Roshni Sahoo, Shengjia Zhao, Alyssa Chen, and Stefano Ermon. 2021. Reliable Decisions with Threshold Calibration. *Advances in Neural Information Processing Systems* 34 (12 2021), 1831–1844.

[66] Sahil Sandhu, Anthony L. Lin, Nathan Brajer, Jessica Sperling, William Ratliff, Armando D. Bedoya, Suresh Balu, Cara O'Brien, and Mark P. Sendak. 2020. Integrating a machine learning system into clinical workflows: Qualitative study. *Journal of Medical Internet Research* 22, 11 (11 2020). https://doi.org/10.2196/22421

[67] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": Supporting clinical decision-making with deep learning. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 99–109. https://doi.org/10.1145/3351095.3372827

[68] Mark Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, Kelly Kester, Sahil Sandhu, Kristin Corey, Nathan Brajer, Christelle Tan, Anthony Lin, Tres Brown, Susan Engelbosch, Kevin Anstrom, Madeleine Clare Elish, Katherine Heller, Rebecca Donohoe, Jason Theiling, Eric Poon, Suresh Balu, Armando Bedoya, and Cara O'Brien. 2020. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics* 8, 7 (7 2020), 1–16. https://doi.org/10.2196/15182

[69] Nigam H. Shah, Arnold Milstein, and Steven C. Bagley, PhD. 2019. Making Machine Learning Models Clinically Useful. *JAMA* 322, 14 (10 2019), 1351. https://doi.org/10.1001/jama.2019.10306

[70] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6

(4 2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118
[71] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (9 2020), 109–124. https://doi.org/10.17705/1thci.00131
[72] James Stewart and Robin Williams. 2005. The Wrong Trousers? Beyond the Design Fallacy: Social Learning and the User.
[73] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter P C Boon, and Ellen H M Moors. 2020. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology* 30, 10 (10 2020), 5525–5532. https://doi.org/10.1007/s00330-020-06946-y
[74] Lucy Suchman, Randall Trigg, and Jeanette Blomberg. 2002. Working artefacts: ethnomethods of the prototype. *The British Journal of Sociology* 53, 2 (06 2002), 163–179. https://doi.org/10.1080/00071310220133287
[75] Amara Tariq, Saptarshi Purkayastha, Geetha Priya Padmanaban, Elizabeth Krupinski, Hari Trivedi, Imon Banerjee, and Judy Wawira Gichoya. 2020. Current Clinical Applications of Artificial Intelligence in Radiology and Their Best Supporting Evidence. *Journal of the American College of Radiology* 17, 11 (11 2020), 1371–1381. https://doi.org/10.1016/j.jacr.2020.08.018
[76] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine Learning in Mental Health. *ACM Transactions on Computer-Human Interaction* 27, 5 (10 2020), 1–53. https://doi.org/10.1145/3398069
[77] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. *ACM Transactions on Computer-Human Interaction* 30, 2 (4 2023), 1–50. https://doi.org/10.1145/3564752
[78] Wen-Juan Tong, Shao-Hong Wu, Mei-Qing Cheng, Hui Huang, Jin-Yu Liang, Chao-Qun Li, Huan-Ling Guo, Dan-Ni He, Yi-Hao Liu, Han Xiao, Hang-Tong Hu, Si-Min Ruan, Ming-De Li, Ming-De Lu, and Wei Wang. 2023. Integration of Artificial Intelligence Decision Aids to Reduce Workload and Enhance Efficiency in Thyroid Nodule Management. *JAMA Network Open* 6, 5 (5 2023), e2313674. https://doi.org/10.1001/jamanetworkopen.2023.13674
[79] Kicky G van Leeuwen, Maarten de Rooij, Steven Schalekamp, Bram van Ginneken, and Matthieu J C M Rutten. 2022. How does artificial intelligence in radiology improve efficiency and health outcomes? *Pediatric radiology* 52, 11 (10 2022), 2087–2093. https://doi.org/10.1007/s00247-021-05114-8
[80] Kicky G van Leeuwen, Steven Schalekamp, Matthieu J C M Rutten, Bram van Ginneken, and Maarten de Rooij. 2021. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European radiology* 31, 6 (6 2021), 3797–3804. https://doi.org/10.1007/s00330-021-07892-z
[81] Simon Tilma Vistisen, Tom Joseph Pollard, Steve Harris, and Simon Meyer Lauritsen. 2022. Artificial intelligence in the clinical setting. *European Journal of Anaesthesiology* 39, 9 (9 2022), 729–732. https://doi.org/10.1097/EJA.0000000000001696
[82] Dakuo Wang, Liuping Wang, and Zhan Zhang. 2021. Brilliant ai doctor in rural clinics: Challenges in ai-powered clinical decision support system deployment. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445432
[83] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2019. ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases. In *Advances in Computer Vision and Pattern Recognition*. 369–392. https://doi.org/10.1007/978-3-030-13969-8{_}18
[84] Thomas Weikert, Joshy Cyriac, Shan Yang, Ivan Nesic, Victor Parmar, and Bram Stieltjes. 2020. A Practical Guide to Artificial Intelligence-Based Image Analysis in Radiology. *Investigative radiology* 55, 1 (1 2020), 1–7. https://doi.org/10.1097/RLI.0000000000000600
[85] Volker Wulf and Björn Golombek. 2001. Direct activation: A concept to encourage tailoring activities. *Behaviour & Information Technology* 20, 4 (1 2001), 249–263. https://doi.org/10.1080/01449290110048016
[86] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Vol. 20. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376807
[87] Wei Xu. 2019. Toward human-centered AI: A Perspective from Human-Computer Interaction. *Interactions* 26, 4 (6 2019), 42–46. https://doi.org/10.1145/3328485
[88] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *DIS 2018 - Proceedings of the 2018 Designing Interactive Systems Conference*. 585–596. https://doi.org/10.1145/3196709.3196730
[89] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fiting intelligent decision support into critical, clinical decision-making processes. In *Conference on Human Factors in Computing Systems - Proceedings (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468
[90] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 4477–4488. https://doi.org/10.1145/2858036.2858373
[91] Nur Yildirim, Hannah Richardson, Maria T Wetscherek, Junaid Bajwa, Joseph Jacob, Mark A Pinnock, Stephen Harris, Daniel Coelho de Castro, Shruthi Bannur, Stephanie L Hyland, et al. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. *arXiv preprint arXiv:2402.14252* (2024).
[92] Hubert D. Zając, Natalia R. Avlona, Finn Kensing, Tariq O. Andersen, and Irina Shklovski. 2023. Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 351–362. https://doi.org/10.1145/3600211.3604766
[93] Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023. Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Transactions on Computer-Human Interaction* 30, 2 (4 2023), 1–39. https://doi.org/10.1145/3582430
[94] Nan-ning Zheng, Zi-yi Liu, Peng-ju Ren, Yong-qiang Ma, Shi-tao Chen, Si-yu Yu, Jian-ru Xue, Ba-dong Chen, and Fei-yue Wang. 2017. Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering* 18, 2 (2 2017), 153–179. https://doi.org/10.1631/FITEE.1700053
[95] Wenying Zhou, Zejun Ye, Guangliang Huang, Xiaoer Zhang, Ming Xu, Baoxian Liu, Bowen Zhuang, Zijian Tang, Shan Wang, Dan Chen, Yunxiang Pan, Xiaoyan Xie, Ruixuan Wang, and Luyao Zhou. 2024. Interpretable artificial intelligence-based app assists inexperienced radiologists in diagnosing biliary atresia from sonographic gallbladder images. *BMC Medicine* 22, 1 (1 2024), 29. https://doi.org/10.1186/s12916-024-03247-9
[96] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>San Jose</city>, <state>California</state>, <country>USA</country>, </conf-loc>) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704