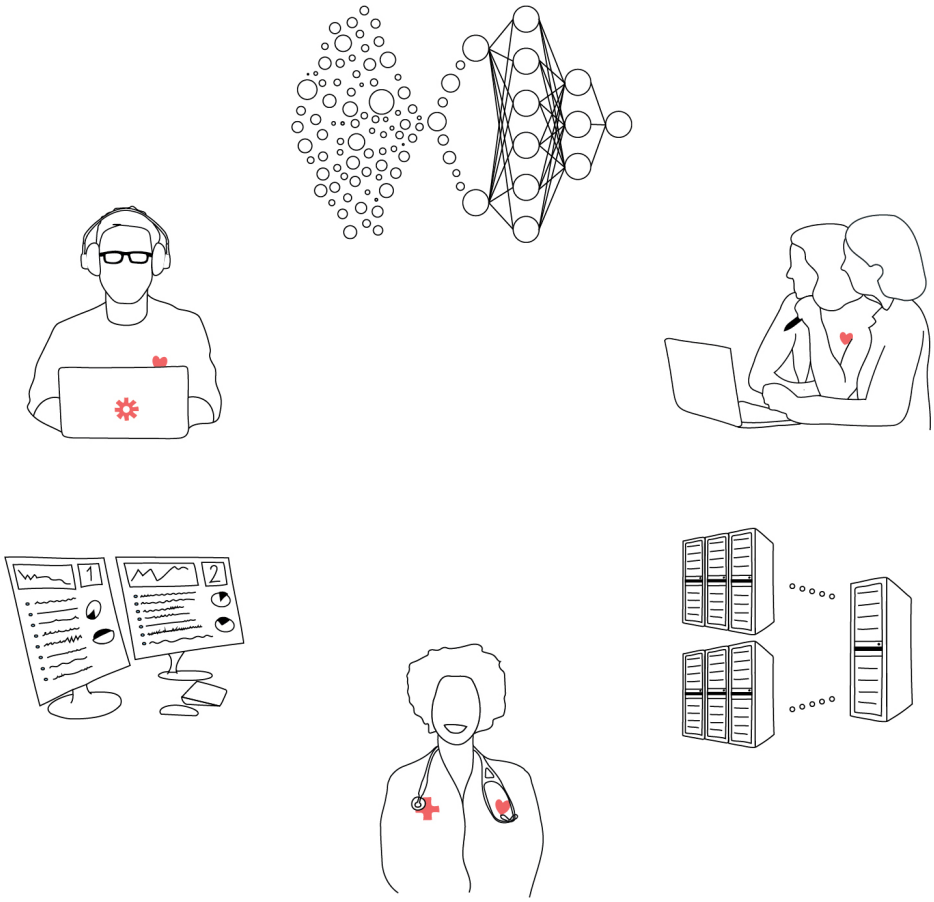
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PHD THESIS

HUBERT DARIUSZ ZAJĄC



IT TAKES A VILLAGE TO RAISE CLINICAL AI

Towards clinical usefulness of AI in healthcare

Supervisor: Tariq Osman Andersen
Co-supervisor: Finn Kensing

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen, February 29, 2024

IT TAKES A VILLAGE TO RAISE CLINICAL AI

A PhD Thesis by

HUBERT DARIUSZ ZAJĄC

Main supervisor

Tariq Osman Andersen
Associate Professor
University of Copenhagen

Co-supervisor

Finn Kensing
Professor
University of Copenhagen

Assessment committee

Committee chair

Naja Holten Møller
Associate Professor
University of Copenhagen

Margot Brereton
Professor
Queensland University of Technology

Niels Van Berkel
Associate Professor
Aalborg University

Department of Computer Science
University of Copenhagen

February 2024

Dziadkom
Halinie i Józefowi Olszewskim

ABSTRACT

Artificial Intelligence (AI) in healthcare, especially based on Machine Learning (ML) techniques, holds significant promise for the field. These techniques have been extensively applied to address various clinical challenges, including pathology detection in X-rays, CT and MRI scans, mammography, skin cancer detection, diabetic retinopathy identification, and predicting readmissions and post-surgery complications. However, despite these advancements, AI-based systems remain notably absent in current clinical practice, limiting their clinical impact. One key factor contributing to this gap is the prevalent technology-centric approach to AI innovation, which often results in the limited clinical usefulness of AI-based support systems.

Through this thesis, comprising this kappa and four of my publications, I address the problem of innovating, i.e., designing, developing, and integrating AI-based systems considered useful by medical professionals in practice. The research presented in this thesis was conducted within the framework of the AI4XRAY project (2020-2025) - an interdisciplinary project aimed at creating a chest X-ray support tool for radiologists in both Denmark and Kenya.

I used a combination of literature review, ethnographic work, and design work to investigate the clinical usefulness of AI-based systems in healthcare. The literature review aimed to identify the challenges of realising AI in clinical practice, while the ethnographic work involved in-situ observations and interviews with medical professionals and AI engineers in Denmark and Kenya. The design work consisted of grounded envisioning and design interventions to explore the opportunities for AI support in chest X-ray practice and configurations affordances of AI support for chest X-ray practice. I analysed the collected data using grounded theory and thematic analysis methods.

This thesis presents five key contributions that contribute to the understanding of clinical usefulness and inform the realisation of clinically useful AI-based systems. Moreover, this thesis emphasises the interdisciplinary nature of clinical AI innovation, making it relevant to practitioners and researchers in Human-Computer Interaction (HCI), AI, and healthcare domains.

First, I enrich the conceptualisation of clinical usefulness with four novel perspectives. I demonstrate how the end-users' expectation for real-world performance depends on the intended use. Additionally, I show how the pre-labelling work on medical datasets for training AI conditions real-world performance, organisational acceptance, and clinical efficacy. Moreover, I highlight how broadening the AI design space enables organisational acceptance and clinical efficacy. I explore how configurable AI boosts organisational acceptance and clinical efficacy. Importantly, these dependencies are not exhaustive, and further research may expand them.

Second, based on a systematic literature review, I find that challenges afflicting the realisation of clinical AI in practice stem not from a single issue but rather from sociotechnical interdependencies present when introducing AI into a clinical context. I conceptualise five challenges spanning three technical (training data & ML model, system integration & data used, and the user interface) and three social (user & system use, workflow & organisation, and healthcare institution & political arenas) aspects. I argue that addressing these challenges necessitates close collaboration among stakeholders with expertise in HCI, AI, and healthcare throughout the innovation processes.

Third, I underscore the importance of attending to the pre-labelling phase in dataset creation. Particularly, I highlight how external and internal factors: regulatory constraints, the context of creation and use, commercial and operational pressures, epistemic differences, and limits of labelling condition the type of data that could be collected, the purpose for which it could be used, and the design of the ground truth schemas, i.e., the selection of labels and additional metrics annotated on the collected data. These fundamental decisions have consequences for shaping the design space of future AI-based systems that use such datasets.

Fourth, I propose five visions for AI support grounded in practical challenges of chest X-ray practice faced across clinical contexts. The visions include distributing examinations by user's expertise, detecting medical emergencies, providing decision support on subtle and difficult cases, measuring visual features and comparing changes across historical examinations, and double-checking reports against radiographs for missed or misinterpreted findings. These visions transcend functionalities traditionally emerging from technology-centred innovation processes and offer nuanced insights into potential AI applications in radiology.

Finally, I delineate how AI-based systems should be configured both before and in use to realise previous visions in practice. The purpose of the configuration is to align the technical dimensions of AI-based systems with clinical needs that depend on social dimensions of clinical practice. The social dimensions span medical knowledge, clinic type, user expertise level, patient context, and user situation. The technical dimensions of AI comprise medical focus, functionality, decision threshold, and explainability methods. By ensuring alignment between these dimensions, AI-based systems can deliver value in concrete situations for concrete medical professionals in clinical practice. I advocate for ongoing consideration of these dependencies to ensure that the AI-based systems undergo necessary configuration before use and include necessary configurability options in use.

SAMMENFATNING

Kunstig intelligens (AI) i sundhedssektoren rummer betydelige muligheder især ved anvendelse af maskinlæringsteknikker. Disse teknikker finder omfattende anvendelse for at imødekomme forskellige kliniske udfordringer, herunder detektion af patologi på røntgen-, CT- og MR-scanninger indenfor mammografi, hudkræft, og diabetisk øjensygdom men også til forudsigelse af genindlæggelser og postkirurgiske komplikationer. På trods af disse fremskridt forbliver AI-baserede systemer bemærkelsesværdigt fraværende i klinisk praksis, hvilket begrænser den ønskede effekt. En nøglefaktor, der bidrager til denne kløft, er den udbredte teknologicentrerede tilgang til AI-innovation, hvilket ofte resulterer i begrænset klinisk anvendelighed ved AI-baserede systemer.

Gennem ph.d.-afhandlingen, som udgøres af denne kappa og fire publikationer, adresserer jeg problemet med AI-innovation, dvs. design, udvikling og implementering af AI-baserede systemer, der betragtes som anvendelige for sundhedsprofessionnelle i hverdagspraksis. Den forskning, der præsenteres i denne afhandling, blev udført inden for rammerne af AI4XRAY-projektet (2020-2025), som er et tværfagligt projekt, med det formål at skabe et AI-baseret beslutningsstøtteværktøj til radiologer både i Danmark og Kenya.

Undersøgelserne anvender en kombination af litteraturstudier, etnografisk feltarbejde samt designmetoder for at undersøge den kliniske anvendelighed af AI-baserede systemer i sundhedssektoren. Litteraturstudiet havde til formål at identificere udfordringerne ved at realisere AI i klinisk praksis, mens det etnografiske feltarbejde involverede situerede observationer og interviews med sundhedsprofessionelle og AI-ingeniører i Danmark og Kenya. Designarbejdet bestod i at udvikle og forankre visioner gennem designinterventioner for at udforske mulighederne med AI-understøttelse samt tilpasning til thorax-røntgenpraksis gennem konfigurationsmuligheder. Jeg analyserede de indsamlede data ved hjælp af grounded theory og tematisk analyse.

Denne ph.d.-afhandling præsenterer fem centrale bidrag, der udvider forståelsen af klinisk anvendelighed ved AI-baserede systemer, samt informerer til realiseringen heraf. Desuden udfolder denne afhandling den tværfaglige karakter af klinisk AI-innovation, hvilket gør det relevant for praktikere og forskere inden for Human-Computer Interaction (HCI), AI og sundhedsområdet.

Først bidrager jeg til udviklingen af begrebet klinisk anvendelighed ('clinical usefulness' på engelsk) gennem fire nye perspektiver. Jeg demonstrerer, hvordan forventningen om AI-systemers ydeevne i den virkelige verden afhænger af det tilsigtede formål, kendt som 'intended use' på engelsk. Dernæst viser jeg, hvordan forarbejdet med opmærkning af medicinske datasæt til træning af AI betinger systemernes ydeevne i den virkelige verden, organisatorisk accept

og klinisk effekt. Desuden fremhæver jeg, hvordan en udvidelse af designrummet for AI samt dens konfigurerbarhed muliggør organisatorisk accept og klinisk effekt.

For det andet gennemfører jeg et systematisk litteraturstudie, som uddyber udfordringerne ved at realisere klinisk AI i praksis og viser, at problemet består af sociotekniske afhængigheder, der udspringer fra introduktionen af AI i den kliniske kontekst. Jeg beskriver særligt fem udfordringer, der spænder over tre tekniske aspekter (træningsdata og ML-model, systemintegration og dataanvendelse samt brugergrænseflade) og tre sociale aspekter (bruger- og systembrug, arbejdsgange og organisation samt sundhedsinstitutionelle og politiske arenaer). Jeg argumenterer for, at håndteringen af disse udfordringer kræver tæt samarbejde mellem interessenter og eksperter inden indenfor HCI, AI og sundhed gennem hele innovationsprocessen.

For det tredje beskriver jeg vigtigheden af at være særlig opmærksom på forhold omkring opmærkning af data i den tidlige fase af AI-innovation, hvilket vedrører dataopsamling og -klargøring. Især fremhæver jeg, hvordan eksterne og interne faktorer (regulatoriske begrænsninger, konteksten for udvikling og brug, kommercielle og operationelle hensyn, epistemiske forskelle og begrænsninger for opmærkning) betinger den type data, der kunne indsamles, formålet, hvormed den kan anvendes, og designet af såkaldte 'ground truth schemas'. Dvs. valget af labels og yderligere metrikker, der er annoteret på de indsamlede data. Disse grundlæggende beslutninger har konsekvenser for udnyttelsen af designrummet for fremtidige AI-baserede systemer, der er baseret på klinisk opmærkede datasæt.

For det fjerde foreslår jeg fem visioner for AI-understøttelse, der er forankret i praktiske udfordringer ved thorax-røntgenpraksis og baseret på forskellige kliniske kontekster. Visionerne inkluderer AI-baseret distribuering af røntgenbilleder efter radiologens ekspertise, AI-baseret detektion af medicinske nødsituationer, AI-baseret beslutningsstøtte ved subtile og vanskelige tilfælde, måling af visuelle funktioner og sammenligning af ændringer på tværs af historiske undersøgelser samt dobbeltkontrol af rapporter mod røntgenbilleder for oversete eller misfortolkede fund. Disse visioner transcenderer funktionaliteter, der traditionelt indgår i teknologifokuserede innovationsprocesser, og tilbyder et nuancerede indblik i potentielle AI-applikationer inden for radiologi.

Endelig gennemgår jeg, hvordan AI-baserede systemer skal konfigureres både før og under brug for at realisere visioner i praksis. Formålet med konfigurationen er at justere de tekniske dimensioner af AI-baserede systemer op mod de kliniske behov, som afhænger af sociale dimensioner af klinisk praksis. De sociale dimensioner omfatter medicinsk viden, kliniktype, bruger-ekspertiseniveau, patientkontekst og brugersituationen. AI's tekniske dimensioner omfatter medicinsk fokus, funktionalitet, beslutningstærskel og forklaringsmetoder. Ved at sikre en god forbindelse mellem disse dimensioner kan AI-baserede systemer levere værdi i konkrete situationer for konkrete medicinske fagfolk i klinisk praksis. Jeg

argumenterer for at overveje disse afhængigheder igennem hele innovationsprocessen for at sikre, at AI-baserede systemer gennemgår nødvendig konfiguration før brug og inkluderer nødvendige konfigurationsmuligheder i brug.

ACKNOWLEDGEMENTS

The breadth and depth of experiences and feelings related to this PhD are quite overwhelming. It is impossible to capture everything, so to start, I would like to thank everyone I met, worked with, sat next to, thought about, talked to, and laughed with because you were part of this journey, and I am so happy where it led to.

Suppose I were to call out a few people. First, it would be the two people who helped me shape my PhD. Tariq and Finn, thank you for your guidance and support. Tariq, I admire how bright you are both in your mind and your heart. It seems as if nothing ever gets you down, and I appreciate that because the PhD sometimes does want to get you down. Finn, I do not think I exaggerate if I say you are the role model for all the PhD students I know. The image of you enjoying life in Coimbra with a glass of wine in one minute and asking the most profound questions about practice at the ECSCW 2024 conference in another minute will be forever with me. I aspire to live like this.

The PhD journey was also about the day-to-day academic life, which for me took place in the Confronting Data Co-Lab. Kristin, Trine, Tina, and Asbjørn, thank you for the laughs, conversations, gossip, drama, and support and for sharing apartments when travelling for conferences and summer schools to save money. Working on the PhD with you made it *objectively* better.

I also want to thank my friends, colleagues, and co-authors, especially Natalia Rozalia Avlona and Irina Shklovski, for their contributions. Irina, if it were not for your ability to make connections (interpersonal and conceptual), the "ground truth or dare paper" would not be even half as good. And speaking of that other half, thank you, Natalia, for writing with me; we produced a solid piece of research that I am very proud of. It is funny how we can find friendships in the darkest of moments (yes, that means writing scientific articles).

I also stayed for half a year as part of my PhD at the University of California, Irvine. Thank you, Yunan Chen, for letting me crash your lab. It was an amazing time that resulted, in my opinion, in an amazing manuscript. I learned a lot from you, and your work ethic is something I aspire to bring forward with me. Also, what could have been a formal PhD requirement turned out to be one of its core parts. This exchange reinvigorated me as a researcher and as a person, and it would not have been the same without the wonderful people I met and who I am happy to call friends, especially Zhaoyuan "Nick" Su and Roman Balakin, Maya Gupta, Bruna de Souza Oewel, David Anderson and Mengleang Ty You.

I want to thank everyone who participated in my studies, the members of the AI4XRAY project who made it possible, but also, or maybe especially, doctors and radiographers from Kenya and Denmark not affiliated with this project. Without your time and effort, there would be no PhD and no point in my work at all.

Finally, I would like to include a few personal mentions. After all, it has been over three years of my life, sometimes exceeding the standard working hours.

To any friends who managed to stick around for the periods of "crunch", "deadlines", and "conferences". I am proud of you. I really appreciate it, thank you. You kept me sane, especially the Morsøgay people: Carlo, Lucy, and Fernando.

I could also not have done it without you, Jens. If I believed in everything I did half as much as you believed in me during this PhD, I would have finished a long time ago. Your support means the world. Thank you!

A na koniec, niespodzianka – kilka słów po polsku. Wszystkim, którzy mnie wspierali w tym długim, dziwnym, trudnym, ekscytującym, kształtującym i nowym procesie, dziękuję. Nie udałoby mi się to bez Was. Szczególne podziękowania kieruję do mojej mamy, taty, brata, dziadków i cichoszów. Ponadto, jestem niesamowicie szczęśliwy, że pomimo tylu trudności w synchronizacji kalendarzy, nadal jesteśmy blisko i nadal udaje nam się organizować wycieczki. Olu, Aniu, Kingo, Moniko, Joanno, te wyjazdy, jak i Wy, dodawały mi sił.

I would also like to thank Augustinus Fonden, William Demat Fonden, and Innovation Fund Denmark for supporting me in this journey. As well as André Miede for developing the *classicthesis* LaTeXtemplate used for this thesis.

PUBLICATIONS

INLCUDED IN THIS THESIS

Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. "Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI." In: *ACM Transactions on Computer-Human Interaction*. Vol. 30. 2. Apr. 2023. DOI: 10.1145/3582430.

Hubert D. Zając, Natalia R. Avlona, Finn Kensing, Tariq O. Andersen, and Irina Shklovski. "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*. Aug. 2023, pp. 351–362. DOI: 10.1145/3600211.3604766.

Hubert D. Zając, Tariq O. Andersen, Elijah Kwasa, Ruth Wanjohi, Mary K. Onyinkwa, Edward K. Mwaniki, Samuel N. Gitau, Shawnim S. Yaseen, Jonathan F. Carlsen, Marco Fraccaro, Michael B. Nielsen, and Yunan Chen. "Towards Clinically Useful AI: Grounding AI Visions in Radiology Practices in Global South and North." Submitted February 2024 to ACM Transactions on Computer-Human Interaction.

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. "'It depends...': Configuring AI to Improve Clinical Usefulness Across Contexts." Submitted February 2024 to ACM SIGCHI Conference on Designing Interactive Systems 2024.

NOT INCLUDED IN THIS THESIS

Hubert Dariusz Zając and Finn Kensing. *Mitigating issues in Healthcare AI projects*. CHI'21 Workshop: Realizing AI in Healthcare: Challenges Appearing in the Wild. May 2021.

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. "The added effect of artificial intelligence on physicians' performance in detecting thoracic pathologies on CT and chest X-ray: A systematic review." In: *Diagnostics* 11.12 (Nov. 2021), p. 2206. DOI: 10.3390/diagnostics11122206.

Hubert D. Zając. "Doctoral Colloquium: It takes a village to raise an AI system - realising AI potential in healthcare." In: *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. 2022. DOI: 10.48340/ecscw2022_dc02.

Hubert D. Zając. "Poster: Designing ground truth for Machine Learning - conceptualisation of a collaborative design process between medical professionals and data scientists." In: *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. 2022. DOI: 10.48340/ecscw2022_p04.

Amelia Jiménez-Sánchez, Natalia-Rozalia Avlona, Dovile Juodelyte, Théo Sourget, Caroline Vang-Larsen, Hubert Dariusz Zając, and Veronika Cheplygina. "Towards actionability for open medical imaging datasets: lessons from community-contributed platforms for data management and stewardship." In: *arXiv preprint arXiv:2402.06353* (Jan. 2024).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

TOWARDS CLINICAL USEFULNESS OF AI IN HEALTHCARE

1

A DREAM THAT HAS NOT YET COME TRUE

Artificial Intelligence (AI) in healthcare, particularly based on Machine Learning (ML) techniques, has promised us the moon - easier and faster access to healthcare, lower cost, higher quality of care, and increased job satisfaction of medical professionals [48]. There have been good reasons to believe in the imminent advent of AI-powered healthcare delivering on these promises. AI techniques have been applied to a multitude of problems across healthcare domains [137, 325]. The academic community have been consistently outputting new cutting-edge AI models that were able to increase detection rates of pathologies on X-rays and CT scans [210], aid breast [126], brain [176] or skin [324] cancer detection, reduce diagnostic and therapeutic errors [97, 130, 204, 259], identify arrhythmia on electrocardiograms [16], or supporting polyp detection [337]. However, when looking at current clinical practice, AI-based systems are vastly absent, and their clinical impact is limited [82, 184].

One of the reasons for the hitherto failure in practice is the technology-centric paradigm of AI innovation [43, 71, 131, 299, 304, 305, 349, 365]. It is important to note that, aware of other connotations of the word innovate, I will use it to refer to the entirety of work on AI-based systems, from conceptual development to clinical integration and routine use. AI innovation, then, is often initiated, shaped, and evaluated with a technology-first outlook [71, 156, 304, 305, 349, 365]. Especially the early data and model work, typically owned by the AI domain [6], are foundational to the capabilities of later AI-based systems [240, 289]. This work defines the domain addressed by the AI model, the ground truth it will use, and the output it will provide. These aspects define the design space of future systems. Moreover, due to their costly nature, altering them past completion may not always be feasible [6, 237, 240].

Despite the technology-centric paradigm, theoretical, methodological, and practical competencies from Human-Computer Interaction (HCI), AI, and health are needed to innovate clinical AI-based systems [42, 43, 240, 326]. However, this collaboration is afflicted by several challenges: lack of mutual understanding, shared terminology, goals, and methods and techniques of work [43, 132, 343, 355], which are further compounded by the difficult of working with AI as a medium [217, 238, 253, 329]. For example, AI-based systems require large amounts of high-quality training data [65, 341]. Designing and prototyping depend on AI capabilities [104], which tend to take final shape only after deployment [355]. Obtaining feedback at the early stages of development, especially in clinical settings, is challenging [77, 356]. As a result, AI-based innovation projects are inherently interdisciplinary and collaborative endeavours underlined by complex and interdependent technology [354].

In such interlinked projects where the work on models, data, and performance is the primary focus, the human and social aspects may fade into the background. However, failing to address these issues throughout the innovation processes could result in uncertainty regarding the sociotechnical alignment of future systems [169, 281, 325, 330]. Such misalignments may jeopardise the final clinical usefulness of the AI-based systems, even if their technical aspects are exceptional. For example, Beede et al. [26] reported how the differences in available facilities between the clinic of use and the clinic of development led to a failure to assess 20% of data by a previously validated and acclaimed AI model. Similarly, Hollander et al. [155] described a system aimed to support decision-making about ED admission due to heart problems. Due to a misconception of clinical practice, the system relied on data that was available only after a clinician had already made the admission decision, which rendered said support irrelevant. Importantly, the sociotechnical hindrances do not refer only to integration challenges. Lehman et al. [208] conducted a prospective study of an AI-based decision support system for mammography to investigate its effect. This modality is one of the few modalities in radiology that has consistently seen clinical use of AI. However, the use of AI reportedly had no "established benefit to women." These studies show that the clinical usefulness of hitherto AI-based support systems is limited and that the predominantly technology-centric innovation process fails to capture the sociotechnical aspects of clinical work the systems are supposed to contribute to. Or, to quote authors of a paper describing a positive retrospective study of a clinical AI-based system, "the usefulness of the proposed... system[s] in clinical practice is still unknown" [12].

2

WHAT TO EXPECT FROM THIS THESIS?

The problems of realising clinical AI have many origins. Solving all of them is not my goal and extends far beyond the content of this thesis. However, these problems lead to the same conclusion - AI-based systems do not provide enough value in clinical practice. This PhD was completed as part of a broader innovation project, AI4XRAY. Within this project, I conducted four studies described in four papers that, together with this kappa, comprise my thesis. The overarching goal of this thesis is *to improve the understanding of the clinical usefulness of AI-based systems.* I will do this by presenting four concise contributions that may be used by interdisciplinary teams to support the realisation of clinically useful AI-based systems. Moreover, each of the studies deepens the understanding of clinical usefulness. The following four questions frame the contributions:

RQ1: What are the challenges of realising AI in clinical practice?

RQ2: What factors condition the creation of training data?

RQ3: What are the opportunities for AI support in chest X-ray practice?

RQ4: How to configure AI-based systems for clinical usefulness across clinical contexts?

In the following sections, I will provide background information about the AI4XRAY project and its influence on this thesis (Section 3), introduce and motivate the choice of research methods used throughout the thesis (Section 4), and unpack the concept of clinical usefulness (Section 5). Next, I will summarise my four research papers, including four contributions towards understanding clinical usefulness related to the intended use of AI, medical datasets, design space, and AI configurability. Alternatively, readers may engage with the full papers in this thesis's second part for in-depth insights. Finally, I will summarise all the contributions and mark directions for future research.

3

RESEARCH SETTING

This thesis was conducted on the foundation of a larger project to innovate an AI-based chest X-ray support system for radiologists in Denmark and Kenya - AI4XRAY - funded by the Innovation Fund Denmark (0176-00013B). This means that this project was commenced primarily as an innovation project with the aim of integrating the developed system into clinical practice in both countries. All the papers included in this thesis either resulted from my work for the AI4XRAY project or work that aimed at supporting it. As a result, most healthcare-related findings will be grounded in radiology.

AI4XRAY was an interdisciplinary collaboration between the University of Copenhagen, the Copenhagen University Hospital (Rigshospitalet), and Unumed. Unumed is a Copenhagen-based startup delivering hospital management systems primarily in Kenya and Indonesia. The project group comprised researchers with expertise in HCI, image analysis, and Natural Language Processing (NLP) from the University of Copenhagen; radiologists and radiographers from the Copenhagen University Hospital; and machine learning engineers and business partners from Unumed.

The project group members dictated the choice of Denmark and Kenya as the focal countries. Denmark was selected due to the origin of the funding and the involvement of the capital region through Rigshospitalet. They also granted access to historical data from the capital region to train the AI model. The capital region is the easternmost administrative region of Denmark and is responsible for providing healthcare services in Copenhagen and adjacent municipalities. Kenya was selected because of the strong business presence of Unumed, which, in line with the project proposal, is set to commercialise the innovated AI-based system.

3.1 BACKGROUND INFORMATION ABOUT X-RAYS

First, X-rays, or chest radiographs, are the most commonly utilised medical imaging globally. They are used to detect lung conditions, evaluate heart conditions and chest injuries, monitor medical devices, and conduct screening and pre and post-operative assessments. On top of the wide range of uses, chest radiographs are cheap, reliable, fast, and expose patients to significantly lower dosages of radiation in comparison to more advanced medical imaging, such as Computed Tomography (CT). As a result, approximately one in every five medical images captured yearly was a chest X-ray, with a total amount of about 2 billion images worldwide [4, 346]. This number is expected to grow as more people gain access to medical care. However, for optimal use of radiographs during the diagnostic process, they need to be interpreted by radiologists. Severe staff shortages worldwide re-

sult in less time available per examination, stress, and overburdening of medical professionals.

Second, chest radiographs are one of the most complex medical images to interpret [124], partially because many conditions share ambiguous visual impressions. Hence, radiologists rely highly on their expertise and experience to disambiguate them, which makes the process exceptionally subjective. As a result, it has been found that there is a high level of disagreement among radiologists when interpreting chest radiographs, with rates reaching as high as 30% [113]. Interestingly, that variability is also present when comparing two interpretations of the same radiograph by the same radiologist at different points in time [211]. The complexity of chest X-rays is also reflected in the number of misclassified and missed findings. It was observed that 3-6% of all chest X-ray studies in the UK included major clinical errors [55, 282], with the ratio growing to 30% for minor ones [61]. The superimposition of internal organs, tissues, fluids, clothing, medical devices, and other objects heightens the risk of overlooking some of the findings. Almost 19% of early lung cancers manifested as nodules on chest X-rays are missed [38].

3.2 DENMARK AND KENYA AS FOCAL COUNTRIES

I will start by briefly introducing both countries and highlighting information relevant to the understanding of the thesis. Denmark is characterised by its well-established welfare state and high-income economy. The Danish state boasts a vast, digitalised public healthcare system. There are 12,000 people per radiologist, and up to 650,000 chest radiographs are captured yearly [153]. Kenya is an emerging economy with diverse natural resources, a growing population, and a colonial history. Kenyan healthcare relies heavily on private practices to cater to the healthcare needs of its populace. There are approximately 265,000 people per radiologist, and there are no statistics about the annual number of captured images [95].

Thanks to the AI4XRAY project, this thesis offers a unique perspective into the innovation work catered to two culturally, economically, and geographically distinct countries. Usually, most AI innovation occurs in the Global North due to the challenge of obtaining access to large amounts of medical data. These data are more readily available in the digitised countries of the Global North, which also tend to host big tech companies and have more resources for research and development. As a result, a disproportionate number of systems are developed to meet the needs of the majority in the Global North. At the same time, the rest of the world often has to adapt or use sub-optimal or biased systems [30, 228, 229]. This poses additional challenges, as on top of challenging local innovation, translating such systems across contexts further exacerbates the complexity [248, 342] (see, e.g., [26]).

HCI scholars suggest that focusing together on commonalities is crucial for counterbalancing the predominant northern viewpoint in

technology design, reducing biases, and facilitating the adaptation of clinical AI-based systems across different regions [168, 201]. Failure to address sociotechnical and political disparities within and between Global North and Global South countries can significantly affect the design and effective use of clinical IT systems [26, 168, 248, 252, 342]. Hence, bridging these gaps is essential to ensure equitable access and usefulness of healthcare technologies worldwide. My last three papers acknowledge these challenges and discrepancies and present work that aims to bridge these gaps and inform innovation of clinical AI-based systems applicable across global lines.

4

RESEARCH METHODS

The methodologies employed throughout the studies align with the progress of the AI4XRAY project and reflect my standpoints as a researcher. I see Participatory Design as my epistemological basis. One may assume this is a natural consequence of having Finn Kensing as a co-supervisor. However, I must surprise those of you who thought so. It was primarily the principles and line of reasoning about IT innovation that made participatory design a sensible choice when dealing with AI innovation (And a little bit of Finn's guidance).

First, Participatory Design emphasises the need to understand the environment designers and developers are going into to conduct IT innovation. The environment spans potential end-users and their work practices, relations, and the goals and purposes of the organisations they work at [187]. Designers and developers must gain this broad understanding through first-hand experience. This is critical in the context of clinical AI, as a poorly and narrowly understood environment leads to low clinical usefulness.

Second, Participatory Design stresses the need for meaningful and broad collaboration that builds on mutual understanding and supports a real sense of agency [187]. This paradigm of working shies from extractive collaboration, where participants or colleagues are only informants in design and development activities. At the same time, this approach does not require everyone to be a designer. Rather, it supports joint sensemaking and ideation about possible futures. Ones that originate in real problems and practices and are not attempts at retrofitting preconceived IT solutions.

These two lessons from Participatory Design can be seen in my choice and use of methods applied to answer the research questions and contribute to the AI4XRAY project (Table 2). I can outline three main types of methods that I used in alignment with the developments in the project and my improving understanding of innovating AI-based systems for radiology: (1) literature review, (2) ethnographic methods, and (3) design methods. While I conducted the systematic review as the first step of my PhD, the ethnographic and design work were often intertwined, informing each other.

Before I describe in detail the methods used, I want to highlight that these methods do not reflect the work necessary to gain access and establish rapport with potential collaborators. Due to the fact that my thesis is interdisciplinary at its core and addresses different clinical contexts, the time and effort necessary to conduct the studies was even longer. In total, I engaged with medical professionals from nine medical sites in Denmark and Kenya (Table 1), access to which was partially initiated by the members of the AI4XRAY project and partially resulted from my outreach efforts. Finally, the first six months of my thesis happened towards the end of the restrictions imposed

#	CLINICAL SITE	TYPE	RADIOLOGISTS	COUNTRY
D1	Copenhagen University Hospital	Specialised hospital	100+	Denmark
D2	Røntgen - Ultralyd Klinikken	Imaging clinic	<5	Denmark
D3	North Zealand Hospital	General hospital	<20	Denmark
D4	Zealand University Hospital	General hospital	<5	Denmark
K1	The Aga Khan University Hospital	Specialised hospital	<20	Kenya
K2	Stratus Medical Imaging Solutions	Imaging & teleradiology clinic	1	Kenya
K3	Adventist Hospital	General hospital	1	Kenya
K4	Coptic Hospital	General hospital	<5	Kenya
K5	The Nairobi Hospital	General hospital	10	Kenya

Table 1: Clinical sites involved in this thesis.

to counter the COVID-19 pandemic, effectively preventing access to clinical sites for nonessential visitors.

4.1 LITERATURE REVIEW

Conducted to answer
RQ1: What are the challenges of realising AI in clinical practice?

I considered learning from past research on clinical AI innovation the best way to locate myself in the field and inform my future studies. This is why I conducted a systematic review to learn about the previous work within the AI, HCI, and health communities. Systematic reviews are invaluable tools for understanding complex issues [98, 99]. Thanks to their structured approach to data exploration, extraction, and synthesis, systematic reviews provide a rigorous and profound overview of the current state of knowledge [108]. However, the issue with this particular task is not the complexity of the studied phenomenon but the diversity of sources, methods, and conceptual frameworks. Rather, HCI, AI, and health researchers are interested in investigating clinical AI innovation from their perspectives. To build a holistic understanding of the challenges faced, I chose to conduct a systematic literature review. Findings from this work were also disseminated within the AI4XRAY project to sensitise the working group about possible challenges ahead.

4.2 ETHNOGRAPHIC METHODS

4.2.1 *In-situ observation*

Conducted to answer
RQ2: What factors condition the creation of training data?
RQ3: What are the opportunities for AI support in chest X-ray practice?

In-situ observations are crucial in informing clinical AI innovation by providing real-world insights into the practical challenges of clinical

RESEARCH QUESTION	METHOD	DETAILS
RQ1		
	Systematic review:	Period: April - October 2021 Papers screened: 9331 Papers eligible: 25
RQ2		
	In-situ observation:	Period: May 2021 - October 2022 Pre-labelling workgroup meetings: 15
	Interview:	Period: May 2021 - May 2022 Participants: 12 Medical professionals & AI engineers
RQ3		
	In-situ observation:	Period: April 2021 - April 2023 Clinical sites: D1, D2, D3, D4 Total duration: 35 hours
		Period: January - February 2023 Clinical sites: K1, K2, K3, K4, K5 Duration: 32 hours
	Interview:	Period: April 2021 - March 2022 Clinical sites: D1, D2, D3, D4 Participants: 20
		Period: January - February 2023 Clinical sites: K1, K2, K3, K4, K5 Participants: 10
	Grounded envisioning:	Period: April 2021 - April 2023 Clinical sites: D1, D2, D3, D4 Participants: 6
		Period: January - February 2023 Clinical sites: K1, K2, K3, K4, K5 Participants: 7
RQ4:		
	Design interventions: *(incl. design sessions)*	Period: January - November 2023 Sessions: 19 Participants: 13 Clinical sites: D1, D2, K1, K2, K3, K4, K5

Table 2: Overview of data collection.

work [118]. Inspired by ethnography, a lived experience of practice provides a profound understanding of the observed problems, processes, and relations [187]. Through observations, researchers can see how the work is really conducted in contrast to learning about it from written or oral reports, which may often distort (consciously or not) the view of the work [318]. This is particularly relevant when investigating matters previously not described in the literature, where my ability to cross-validate reports would be limited. This is why I conducted participatory observations of work conducted by a group of AI4XRAY members before data labelling (15 meetings and workshops between medical professionals and AI engineers). Thanks to first-hand observations of the collaboration, I could capture its essence that otherwise was not reflected in produced artefacts or notes taken by other participants. Namely, the written accounts from the meetings focused on concrete decisions regarding labels and labelling. On the other hand, through the observations, I uncovered the reasons behind the discussions about labels and captured what motivated its participants.

Moreover, unlike work reports, in-situ observations do not rely on the perception of work by other people. To truly learn how the work is done, one must experience it. Only then will the design of systems that aim to alter the practice be meaningful. This is one of the core premises of Computer Supported Cooperative Work (CSCW) [118, 269]. This is why I conducted 67 hours of in-situ observations in nine medical sites in Denmark and Kenya with 22 radiologists and radiographers. This allowed me to gain a deep understanding of radiologists' work in diverse clinical settings. Importantly, this knowledge included more than just the work of interpreting radiological examinations. I could contextualise specific tasks of radiologists within the complex realities of healthcare delivery, including variations in patient populations, clinical practices, and technological infrastructures. I learned about the broader context of clinical work, a collaboration between medical professionals, and the rhythm of the clinics - knowledge necessary to design AI-based systems that fit the real clinical context and not a carved-out and sensitised version of it. I used it to inform the work in the AI4XRAY project and the development of a prototype used in the last study.

4.2.2 *Interview*

Conducted to answer
RQ2: What factors condition the creation of training data?
RQ3: What are the opportunities for AI support in chest X-ray practice?

It is hard to imagine qualitative or ethnographic studies without interviews. This method is essential to give meaning to observed practice, to learn about nuance, and to learn participants' perspectives on investigated issues [187, 197]. For example, interviews provide an

opportunity to explore divergent viewpoints and complexities that may not be apparent through observation or other methods. Moreover, interviews can be used to support mutual learning. Well-defined open-ended questions may spark reflection and learning about topics relevant to AI innovation, resulting in more fruitful collaboration at later stages.

I used interviews throughout the thesis - from interviews with project members to better understand their roles (Paper II); through interviews with physicians ordering chest X-rays to understand the rationale behind using this type of medical imaging (not part of any of my papers); to interviews during in-situ observations to clarify any misunderstandings about observed actions (Paper III). This versatility of contributions was possible thanks to the flexibility of the interview as a method. I followed a format appropriate to the type of information inquired about, i.e., structured questions following a guide when inquired about feedback on a prototype (part of work described in Paper II) to almost conversation-like when discussing nuances of local practice (Paper III).

4.3 DESIGN METHODS

4.3.1 *Grounded envisioning*

Conducted to answer
RQ3: What are the opportunities for AI support in chest X-ray practice?

Grounded envisioning at its core draws from speculative design [17]. While not part of an established method, similar futuring, described under varying names, was successfully used in other AI-focused studies [180, 314, 322]. I used it to investigate how AI-based systems could support radiologists in their practice, which is a crucial question from the perspective of the AI4XRAY project. Due to the low usefulness of previously evaluated AI-based systems and their primarily technology-driven origin, I wanted to explore the space of AI support as viewed by radiologists, focusing on their understanding of valuable support.

To carry it out, I conducted in-situ observations and in relation to any observed doubts, confusion, or encountered problems, I asked radiologists whether they could envision any way that AI could help them in such situations. This approach uncovered ideas transcending the traditionally explored second opinion and screening, e.g., measuring visual features on X-rays or assessing completed reports against detected radiological findings to ensure that radiologists had not missed any. We confirmed that the space of radiological AI support may have been underexplored and that there are opportunities other than a conventional second opinion worth pursuing.

I think two conditions are critical to the successful use of this method. First, almost all radiologists I worked with had previous

exposure to AI-based systems and were disillusioned about AI capabilities, i.e., they did not use AI-based systems in their daily practice, even if such were available. This condition implicitly anchored their visions in or close to the space of current AI capabilities, which is a crucial step of speculation for design [17]. Second, the envisioning was conducted as part of the in-situ observations. Developing visions for and ideas about a system's functionalities in a lab or IT office risks diverging from the actual needs emerging in practice. By addressing real issues in the context of their practice, I wanted to maintain the link between functionality and practice and thus increase the chances of clinical usefulness.

Finally, the visions of support were used to guide the development of a prototype of the AI4XRAY system, which was used to conduct the design interventions. Otherwise, the dissemination of findings from this work was limited due to two factors. First, the publication cycle required additional work to finalise Paper III that contained the results of the grounded envisioning. Second, at the time of writing, the AI4XRAY project suffered independent data labelling delays, which resulted in speculation about narrowing the project scope and focusing on a narrow subset of radiological findings. Due to this direction, visions explored with radiologists may not be applicable within AI4XRAY.

4.3.2 *Design interventions*

Conducted to answer
RQ4: How to configure AI-based systems for clinical usefulness across clinical contexts?

A design intervention, as defined by Halse and Boffi [141], is a method that cuts across design and ethnography, a method of complexification, and a method that "enables new forms of experience, dialogue, and awareness about the problem to emerge" (also [44, 46]). It is a form of inquiry that is experimental and supports positioning "in-between what is already there and what is emerging as a possible future" [9]. The most vivid difference between usability studies or prototyping and design intervention is its open-endedness. Instead of informing concrete design decisions, they are used to explore and direct future work. Design interventions should also occur during clinical practice, i.e., we as designers should intervene in radiologists' everyday work settings. This influences participants to consider solutions and envision alternatives to existing systems, all while being mindful of the specific needs of their local context. In this thesis, we used design interventions to map the dependencies of clinical usefulness rather than seeking numerical or visual answers to concrete usability questions.

As part of the Paper IV, we conducted eleven design interventions. Additionally, we conducted eight design sessions, which followed the same exploratory goals but lacked the interventionist aspect. I will re-

fer to them jointly as design interventions, as this type comprises most of this work and better reflects its goals. In total, we conducted the interventions with thirteen radiologists from seven clinical sites in Denmark and Kenya (avg. 60 minutes). We strove for each intervention to be grounded in participants' work practices to situate their input instead of trying to generalise their practice to undefined contexts.

The interventions used three versions of a prototype of an AI-based support tool to practically explore the topic of usefulness. The prototype relied on an AI model developed within the AI4XRAY project. Consequently, participants engaged with authentic historical data and corresponding real AI predictions. The only clinical difference between the practice and the intervention was the anonymisation of data and the lack of auxiliary clinical information available.

This research approach proved essential for the thesis, as it grounded visions of how AI could be used in clinical settings across different hospitals. Moreover, through iterative changes to the prototype, we explored how configuration affordances can help align the support offered with the local needs. Consequently, we conceptualised concrete recommendations for designers and developers innovating radiological AI-based systems.

Finally, similarly to the results from grounded envisioning, the results of the design interventions have not yet been published. Moreover, these findings have not yet been incorporated within the AI4XRAY project due to its changing focus and delayed development. However, we hope to see them enacted as ensuring usefulness has been named the core priority for the future work of the project group.

4.4 DATA ANALYSIS METHODS

I employed two analysis methods to make sense of the collected qualitative and empirical data: grounded theory [73] and thematic analysis [56].

I followed the grounded theory principles in my first two papers (Paper I, and Paper II). This choice was dictated by their goal of open exploration of a topic. Additionally, the breadth of collected data allowed for building more robust and encompassing conceptual contributions about broader topics, like dependencies of pre-labelling stages of dataset creation or challenges afflicting the real-world realisation of clinical AI. Without previous assumptions about the data content, grounded theory proved to be an appropriate method for rigorous interpretation and analysis.

To analyse data in Paper III and Paper IV, I used thematic analysis. The analysis performed for these two papers differed from the previous ones, as I had gained a deeper understanding of the domain and had better expectations regarding the data. Additionally, the aim of these papers was not to derive high-level theory but to explore in-depth the experiences of medical professionals in their work. This

type of analysis allowed me to understand how radiologists' work practices are connected to opportunities for AI support and how AI-based systems can be configured to provide clinically useful support across medical contexts.

The analysis of data for all the papers was marked by iterative reflection. The only way to know that what I conceptualised was true and valuable was to cross-check it with other data or bring it back to practitioners to discuss it with them. In most cases, I relied on revisiting data to understand better its context of creation and joint sensemaking with my supervisors. In other instances of my work, especially pertaining to artefacts like the AI-based support system prototype and a tool designed for labelling work within the AI4XRAY project, I was able to validate my assumptions with the practitioners.

The problem of when to stop collecting data was solved rather practically. I could not conduct fieldwork indefinitely in clinical sites on different continents, countries, and cities. My participants were busy individuals working in high-stakes environments, so my engagements with them were limited to the prearranged time they felt comfortable committing. In situations when only limited time was possible, I aimed at recruiting participants in similar settings to ensure enough data coverage. Finally, relatively similar work practices of doctors in similar clinical settings (described in Paper III) helped to insights from data collected across settings.

5

WHAT DO WE TALK ABOUT WHEN WE TALK ABOUT THE CLINICAL USEFULNESS OF AI?

In this thesis, clinical usefulness serves as a central concept anchoring the exploration of AI-based systems in clinical practice. Unlike typical medical interventions evaluated through randomised clinical trials or conventional IT systems with established methods of innovation, these systems are intricate sociotechnical systems [109]. They have to dovetail contributions from HCI, AI, and health into a cohesive vision to bring meaningful change to clinical practice [43, 117, 169]. Clinical usefulness encompasses these three domains' diverse goals and values and highlights their de facto common goal. Understanding this interdisciplinary foundation provides essential context for engaging with the specific contributions outlined in the four papers of this thesis.

At this point, it is important to acknowledge other attempts at a more holistic framing of AI-based systems in clinical practice. Kocak et al. [190] investigated the clinical utility of AI models in radiology, understood as offering "improvements on radiologists' clinical evaluation or traditional algorithms". This work made the important distinction between performance evaluation and clinical value. Recently, Boverhof et al. [53] proposed a 7-step evaluation framework of AI-based systems for radiology focusing on clinical value. This framework not only differentiated three distinct ways AI-based systems may contribute to clinical practice (through diagnostic thinking, therapeutic, and patient outcome efficacy) but also pointed to local efficacy as the applicability of a system in the local context. Finally, Blandford and colleagues [42, 43] opened up the clinical outlook on AI to HCI concerns and acknowledged the need for organisational and end-user acceptance. These studies show the change in understanding of the nature of AI-based systems as sociotechnical. I build on this work and attempt to provide the most encompassing definition of clinical usefulness.

I understand clinical usefulness as *the overarching quality of AI-based systems emerging from the interplay of their real-world performance, clinical efficacy, and organisational acceptance in a situated clinical context for specific end-users* (Figure 1). This definition focuses on the high-level contributions from each of the domains. While it does not capture the breadth of considerations attended to within each of the domains, it provides the necessary high-level perspective to consider them in unison. I will explore each component to draw on AI, health, and HCI communities' theoretical, methodological, and practical competencies.

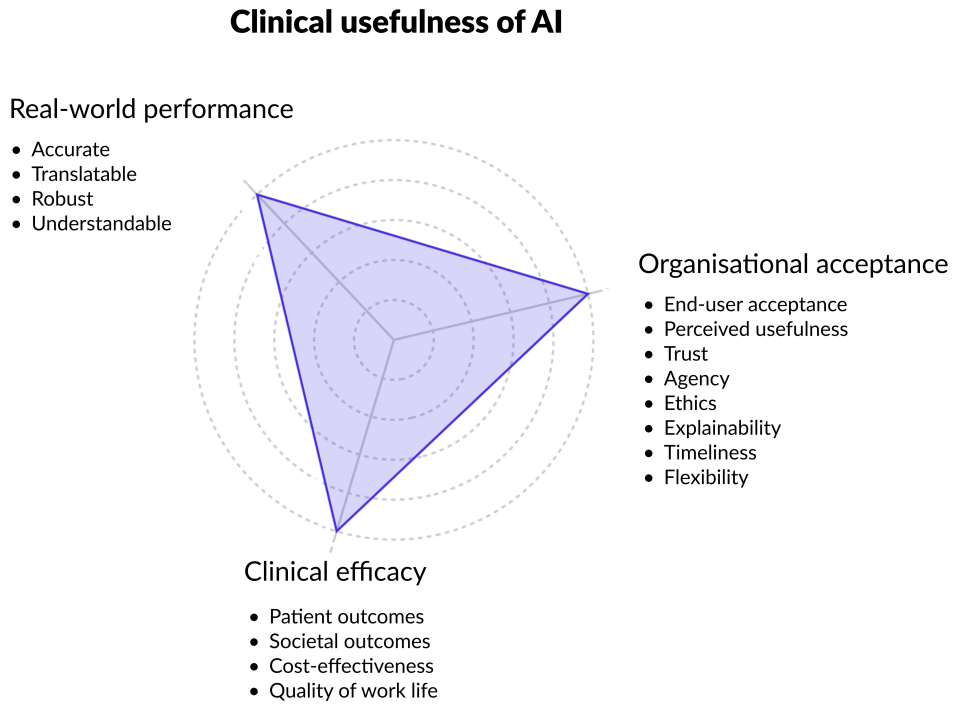


Figure 1: The three components of clinical usefulness.

5.1 REAL-WORLD PERFORMANCE

Clinical usefulness necessitates robust performance. Improvements in AI performance resulted in the renewed interest in applying it across different domains, including health. As such, performance is the cornerstone of clinical AI, and without the ever-improving technology, current innovation would not be possible [291]. Most notably, the arrival of Large Language Models (LLMs) may be considered a technological breakthrough that resulted in a whole new space of opportunities for AI use [60].

However, the performance of AI models in retrospective evaluations often does not hold in real-world settings, jeopardising clinical usefulness. For example, Beede et al. [26] described a peer-reviewed AI model that failed to successfully evaluate 21% of real-world cases during a pilot deployment in a new clinical context. This clear problem with translating AI models across contexts is one of the challenges afflicting performance in the real world. Reddy et al. [275] proposed a framework to guide AI innovation. A special focus on ethics, translation, and generalisability should be applied across the innovation stages to achieve similar performance levels across settings.

The ability to translate models across contexts is often linked to the quality of the training data. Major et al. [220] explored how, in a clinical setting, the selection of training data may improve the correct assessment of performance. In many cases, the data work preceding the model work has a critical influence on final usefulness [250]. For example, certain training data choices may result in AI models utilising learning shortcuts, i.e., learning easily recognisable yet irrelevant correlations that are then used to make general predictions [251]. En-

suring high-quality and representative data supports the final robustness of AI models.

Finally, robust performance does not mean 100% accuracy. Bansal et al. [21] showed how understanding AI capabilities, particularly the reason behind AI's mistakes, supports the effective use of AI predictions. Focusing on how AI can supplement human decision-making with imperfect predictions may induce a greater positive effect than improving absolute accuracy by a few points [20]. Other research has shown that even imperfect algorithms may be useful in practice [52].

5.2 CLINICAL EFFICACY

Clinical usefulness necessitates clinical efficacy [184]. Health researchers increasingly engage with AI not only as recipients and evaluators but also as relevant actors participating in innovation processes [43, 301]. Their understanding of clinical work is crucial, as clinical efficacy is not always a direct consequence of robust performance. [37, 303]. For example, despite successful integration and robust performance, a decision support system for cancer detection on mammography images was found to have no positive effect on clinical outcomes [208]. Such studies, especially in radiology, prompt questions about the clinical value of AI-based systems [206, 313].

In the face of the uncertain clinical value of AI in practice, Blandford et al. [43] suggested closer and more meaningful collaboration between medical professionals, developers, and designers. They argued that AI-based systems fail to provide value in practice due to the lack of understanding and transfer of knowledge across the domains. Recht et al. [272] echoed these suggestions and called for increased AI evaluations in clinical practice to understand opportunities for better support. Importantly, aware of the iterative nature of AI innovation, healthcare researchers have expressed readiness to challenge the traditional evaluation methods, like randomised clinical trials. Particularly, iterative, local, and patient-centred approaches that go beyond technical measures and focus on the quality of care [86, 88, 184, 303].

These propositions foreground change in thinking about the potential value provided by AI-based systems. Traditionally, clinical efficacy has been understood to focus on improved patient outcomes. However, health researchers working with information systems have explored additional ways AI can benefit clinical practice. For example, van Leeuwen et al. [206] proposed a hierarchy of seven types of efficacies that can be used to evaluate AI's impact on clinical practice: technical, potential, diagnostic accuracy, diagnostic thinking, therapeutic, patient outcome, and societal efficacy. Boverhof et al. [53] expanded this model with cost-effectiveness, which traditionally belonged to the assessment of clinical efficacy and local efficacy to account for the translation of models across contexts. Prior to these hierarchies, Bodenheimer and Sinksy proposed the quadruple aim for health system optimisation. Their proposition extended the established Triple Aim: "improving the health of populations, enhancing the patient experi-

ence of care, and reducing the per capita cost". The fourth aim pertained to medical professionals and called for improving their work life. This aim is particularly relevant when considering AI, as research shows that lack of social considerations of work is a common cause for a failure in practice [169, 281, 325, 330].

5.3 ORGANISATIONAL ACCEPTANCE

Clinical usefulness necessitates clinical organisational acceptance, i.e., organisations must see a value in integrating a system, and the end-users must be willing to use it [187]. In the context of clinical AI, the latter has been proven more challenging (see, e.g., [25]). The HCI community explored many avenues to ensure that end users accept and see value in using AI-based systems, highlighting the need to consider the current workflows, work practices, and sociotechnical context [82, 169, 253, 281, 325].

In clinical work, where a single correct answer may not exist, explaining one's reasoning and adhering to standard procedures serves as a strategy to deliver the highest quality of care [105, 110]. Initially, AI functioned in the exact opposite way - provided an answer without a clear indication of the reason. This proved difficult to accept for clinical professionals [159] and sparked questions about ethics [35]. Inadvertently, transparency and explainability are linked with ethics of healthcare delivery, as they help medical professionals make the most appropriate decisions for their patients [166]. Even when an AI-based system does not provide 100% accurate support, through transparency and explanation, even such input may be found useful in clinical practice [21, 52, 221]. HCI researchers supported these clinical decision-making practices by investigating explainable AI (XAI) methods. These have been found to enhance trust and support decision-making and oversight over AI-based systems, fostering the sense of agency in clinical end-users [67, 135, 205, 348].

Another important consideration in healthcare is how the value of information depends on the interplay between its content, time, the place it is shared, and the intended recipient. For AI output to be useful, it has to deliver the right information at the right time and place [263, 357], as both too early [127] and too late [155] deliver jeopardises clinical usefulness. Moreover, researchers point out that a broad range of healthcare workers carry out clinical and care work, often in the shadow of medical doctors [157, 357]. Finding the best way to integrate AI remains an active area of HCI research [169, 290, 302, 357].

Moreover, the position given to AI in the clinical workflow determines how its prediction will be received [179, 356]. This new type of information presentation in a clinical context prompted more HCI research on human-AI collaboration and joint reasoning [34, 36, 52, 68, 69, 100]. Even if not left with the power to make the final call, AI-based systems are often perceived as more capable, effectively am-

plifying their input. This multifaceted authority of AI poses more ethical dilemmas [327].

Finally, HCI researchers point out that there are also mental and temporal costs associated with using AI-based systems. Understanding AI output, assessing its correctness, and considering how to weigh it - all these actions take time and effort. Sometimes, this total cost may be too high to use the AI-based system in daily practice. Cai et al. described how their tool to interact with AI to increase the utility of AI predictions led medical professionals to worry that they spent too much time almost playing with it [67]. On the other hand, a more common worry is the work needed to discern false positive predictions. Researchers reported how excessive work needed to deal with false positive predictions resulted in the failure of clinical AI-based systems in practice [18, 25, 221, 325]. This happens because the benefits of a correct AI prediction are undone by the extra work needed to reap them in a resource-sparse clinical environment.

6

SUMMARY OF PAPER I: CLINICIAN-FACING ML IN THE WILD

Increasingly, we are gaining insight into the challenges of innovating clinical AI-based systems [355]. Researchers are investigating aspects such as trust [267], explainability [348], and design methods [358], or integration [357]. However, these inquiries often focus on exploring specific problems or challenges. Fewer studies explore the broader issues encountered when transitioning from a lab to clinical practice, generally considered the "last mile" problems [82]. Before we committed to the innovation of the AI-based system within the AI4XRAY project, we wanted to understand the complexity of these problems better. Through Paper I, we synthesised knowledge from AI, HCI, and health domains that could help realise ML-based systems in clinical practice. We build on insights from 25 papers describing evaluations of clinician-facing ML-based systems, selected through a three-stage process from 9331 papers. Importantly, due to the vague use of the term artificial intelligence, we included only articles that described machine learning-based systems.

6.1 WHAT ARE THE CHALLENGES OF REALISING ML IN CLINICAL PRACTICE?

An answer to this question is twofold. The first part pertains to the process of clinical ML innovation. The second part addresses the challenges of an ML-based system in clinical practice.

6.1.1 *The challenges of ML innovation process*

If the process of innovating clinical ML is flawed, the final clinical usefulness of created systems may deteriorate. As a relatively new area, the innovation of clinical ML has not yet developed established methods to guide the innovation processes. This was observed in the reviewed studies. The reported methods were diverse, and many were introduced without proper description. The authors resorted to phrases like "collaborated," which further complicated the replicability of reported processes.

We discussed that these challenges were exacerbated by the need for close collaboration of participants from different domains (ML, HCI, and health), as they all represent unique epistemological standpoints, goals, and terminologies. Particularly, HCI practitioners value user involvement and iterative design [132], while ML relies on data-driven experimentation and optimisation [6], and Health emphasises evidence-based practice and clinical trials [70].

We proposed two approaches for the HCI community based on previous research into difficult collaborations on how to address this challenge. First, we suggested joint work among stakeholders on the "demystification" of clinical ML. This means focusing on collaborative design methods that help translate the theoretical ML-focused design decisions to clinical practice frame of reference (see, e.g., [358]). We hope such joint demystification could increase the agency of involved participants and lead to more meaningful collaboration. Second, we advocated for establishing new HCI techniques to address the difficulties of working with ML as a medium. Such new techniques could take place in a "third space" [241], that is, a space not "owned" by any of the domains and should function as a means to discuss and negotiate design, exchange perspectives, terminologies, and ultimately serve as a mutual learning opportunity for all parties involved [50]. This suggestion was motivated by the relative failure of past methods and an organic move towards new propositions, e.g., seen in the "undefined" grounded in "collaboration" methods reported in the reviewed papers. These two suggestions aim to narrow the epistemological and methodological gap between domains that need to collaborate to innovate clinically useful ML-based systems. This need was well captured by Blandford et al. [43], *"Until there is much greater mutual understanding and mutual valuing of the complementary research traditions than exists at present, people risk disappointment and rejection in trying to bridge the divide"*.

6.1.2 *The challenges of clinician-facing ML-based systems in practice*

We identified five types of challenges afflicting the realisation of clinical ML-based systems, which originated as interdependencies between the social aspects of the use and clinical environment and the technical aspects of the systems (Figure 2). The identified *technical* aspects included (1) training data & ML model, (2) system integration & data used, and (3) user interface. The *social* aspects spanned (1) user & system use, (2) workflow & organisation, and (3) health institution & political arenas.

Training Data & ML Model and User & System Use: The quality and quantity of the training data affected the performance and accuracy of the ML model, which in turn influenced users' trust and adoption of the system. Moreover, the choice of training data was linked to how well an ML-based system models clinical practice, affecting its usefulness.

System Integration & Used Data and Clinical Workflow & Organisation: Introducing an ML-based system in clinical practice can disrupt the existing clinical workflows and require changes in roles, responsibilities, and routines. Further, the quality of technical integration into existing systems affected the usefulness of the systems, e.g., integration requiring manual data entry or resulting in delays in delivering ML prediction reduced the clinical usefulness.

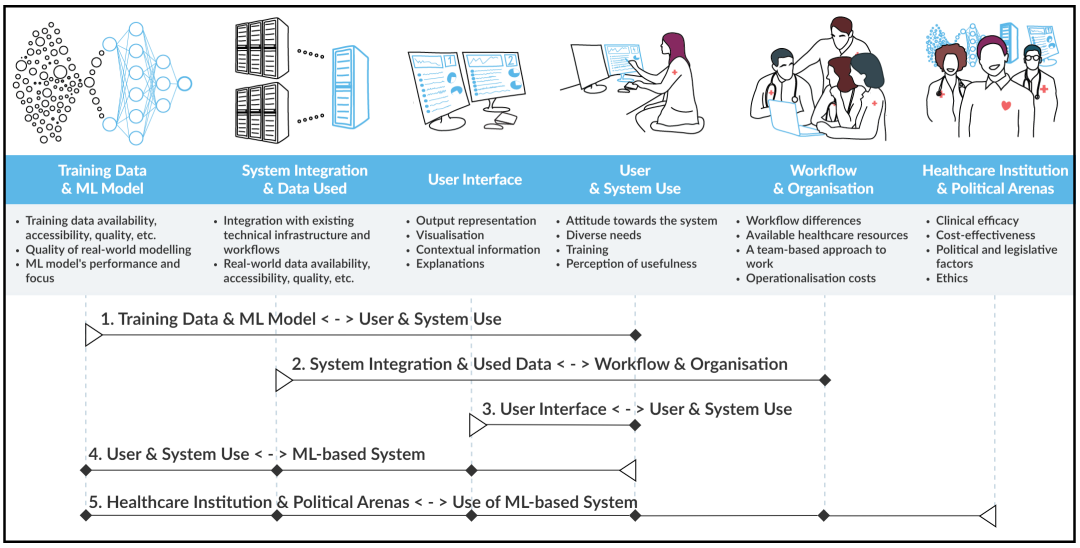


Figure 2: From [362]: Five sociotechnical interdependencies of clinician-facing ML-based system innovation.

User Interface and User & System Use: The user interface design affected how users perceived and interacted with the systems. Explainability, interpretability, and transparency especially affected decision-making practices and the professional agency of medical professionals. New types of information associated with ML output prompted questions of bias and ethical use. The type of explanations influenced the

User & System Use and ML-based System: Clinicians' attitudes significantly shaped their acceptance of ML-based systems, with scepticism and resistance hindering adoption. Additionally, diverse clinical roles and needs led to varied preferences for system functionality, e.g. when using the same system, nurses favoured actionable suggestions while physicians sought dynamic output they could engage with [22]. Finally, barriers such as insufficient end-user training and low computer literacy further impeded successful use in practice.

Healthcare Institution & Political Arenas and Use of ML-based Systems: The innovation of a clinical ML-based system can be influenced by wider institutional and political factors, such as regulations, policies, standards, and incentives. The type of healthcare system and its incentives affected how actionable ML outputs were, thus affecting their clinical usefulness. Further, in medical cases where the diagnosis and treatment are less defined, clinicians may find it more difficult to incorporate another variable into an already ill-defined process. Lastly, ethical considerations invoked by ML outputs were found to affect the perceived usefulness.

We proposed three recommendations for addressing these challenges during ML innovation processes. First, there is a pressing need to adopt an approach of iterative co-configuration, recognising that the successful integration of ML technologies into clinical practice requires ongoing adjustment and alignment of sociotechnical components, including data, ML models, user interfaces, clinical workflows, and organisational goals. This includes viewing the innovation

process as one of organisational change and continuous collaboration. Second, we advocate conducting near-live and real-world experimentation with prototypes and early versions of ML-based systems within clinical settings. Such engagement provides invaluable insights into the practical challenges and opportunities for design, development, and deployment, facilitating meaningful adjustments that enhance system usefulness. Third, design, development, and evaluation activities should be merged into an iterative and cyclical process, ensuring that clinical practice's evolving needs and complexities are addressed throughout the innovation life cycle. By embracing these strategies, coupled with ethical considerations and regulatory oversight, the innovation of clinician-facing ML-based systems can progress toward delivering high-quality care and improving patient outcomes effectively.

6.2 CLINICAL USEFULNESS: THE EXPECTATION OF REAL-WORLD PERFORMANCE DEPENDS ON THE INTENDED USE

In this paper, we used the intended use concept to tease out the different purposes and goals of ML in clinical practice. Based on the reviewed systems, we conceptualised the following intended uses: decision support, prioritisation, and automation. Decision support ML-based systems aid healthcare professionals in making informed decisions for individual patients. Prioritisation ML-based systems operate in the context of multiple patients. They conduct an individual assessment of all relevant patients based on predefined criteria and highlight patients who need providers' attention. Finally, when machine learning-based systems are given autonomous agency and authority, they may replace medical providers instead of supporting them in clinical tasks during healthcare delivery.

We discovered that different intended uses altered clinician's expectations towards the systems. Clinicians using decision support systems prioritised explainability over perfect performance. Factors such as contributing factors, raw input data, and historical context were emphasised, enabling clinicians to interpret predictions effectively and enhancing their clinical skills with a new outlook. When ML was utilised for prioritisation, attention was directed towards positive predictive value, even at the expense of explainability or other technical metrics. The high positive predictive value was considered crucial to reducing false positives, which demanded additional work and resources from the staff, clinic, and patients. As an intended use, automation showed the highest reliance on predictive performance, as it involved AI making the final decision.

I recommend establishing a clearly defined intended use to guide the innovation of ML-based systems from the earliest stages. Such guidance could facilitate better alignment between the technical aspects, user interface and explainability affordances, and the expectations of clinical end-users, resulting in improved clinical usefulness.

7

SUMMARY OF PAPER II: GROUND TRUTH OR DARE

The development of clinical AI conventionally starts either from data or model work. In the AI context, data work is a general term for preparing data for AI development. For supervised training, this work usually spans four stages: data collection, data processing, design of ground truth schema, and data labelling, as conceptualised in Paper II. The downstream effects of this work, beyond the delivered artefacts, are often poorly understood, as data work used to be considered of lesser quality and importance [289].

Critical data scholars and HCI researchers have started in recent years to deconstruct data labelling as the moment in the data workflow where a range of workers use their knowledge and intellect to transform data into a format usable by AI models. This great work described how politics, power, and views are embedded in data through the act of labelling [228, 229, 231]. Even labels applied to medical data by subject matter experts result from a situated design process [240]. The quality of these labels has been shown to affect the performance of clinical AI models [139, 170, 234]. The relationship between the quality of training data and the final AI-based system may be more profound than just affecting AI's performance. Oakden-Rayner et al. found that AI trained on incompletely labelled medical datasets may lead to negative clinical outcomes [251]. This is why we investigated pre-labelling data work and its influence on the final shape of medical datasets.

7.1 WHAT FACTORS CONDITION THE CREATION OF TRAINING DATA?

We analysed data collected during fieldwork in three European health tech organisations (ORG I, II, and III). I collected data within the AI4XRAY project (Org I in the paper). Natalia Rozalia Avlona collected data from the other two organisations. During the course of the AI4XRAY project, the pre-labelling phase spanned nearly a year, during which I was deeply involved from its inception (also described in Paper 5). Initially, I joined as an observer of the collaborative efforts between ML engineers and radiologists, occasionally contributing to the discussions with insights from fieldwork in other hospitals (described in Paper III). As the project progressed, my responsibilities evolved, and I actively contributed to the dataset creation by designing and evaluating a specialised labelling tool in collaboration with radiologists. The work was finalised using the tool to label data to train the AI4XRAY AI model.

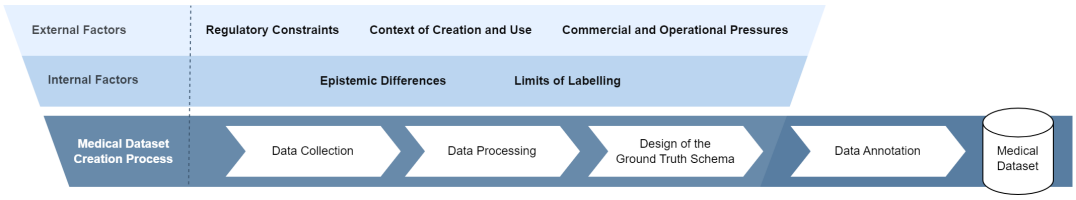


Figure 3: From [361]: A simplified medical dataset creation process expanded with the design of ground truth schema and factors conditioning the pre-annotation stages.

7.1.1 *The Design of Ground Truth Schema*

It is necessary to delineate the activities involved in the process to discuss factors affecting the creation of medical datasets for AI purposes. In this paper, we conceptualised a previously omitted stage of data work - designing a ground truth schema, which we defined as "a collection of relational labels and metrics, as well as their definitions and examples that are used during data labelling" [361]. This critical stage was observed within the AI4XRAY project but also reminiscent in ORG II and III. This work aimed to decide what medical conditions should be labelled, what additional metrics should be labelled, and how the labelling should be conducted. All of these decisions were motivated by creating high-quality data that can be used to train clinically useful AI models within the context of each respective organisation. All of them were also influenced by the five factors conceptualised (Figure 3).

7.1.2 *External Factors*

Three external factors (Table 3) play a role in determining the space for creating medical datasets and what data can be collected and made available for labelling. Thereby, they impact the ultimate structure of the datasets even before the start of labelling.

Regulatory constraints: These are laws (e.g., EU's General Data Protection Regulation or US's Health Insurance Portability and Accountability Act) and standards (ISO 2700013001, and Good Medical or Good Manufacturing Practices) that specify what data can be collected and for what purpose it can be used. They ensure that data collection and processing respect the patients' privacy, consent, and rights. However, they have the broadest consequences for the final AI-based systems. These regulations determine the extent and purpose of data collection, which has downstream effects on every aspect of AI innovation.

Context of creation and use: These are the geographic, demographic, and linguistic aspects that shape both the data (sourced from specific hospitals and specific populations) and the ground truth schema (the choice and meaning of labels). The labelled data has inherent characteristics of the local population, which may not correspond to the populations from the intended country of use. Moreover, medical professionals designing the ground truth schemas defined the labels

REGULATORY CONSTRAINTS
Extent of Collected Data
Predetermination of Purpose
CONTEXT OF CREATION AND USE
Geographic context of use
Demographic context of production
Linguistic context
COMMERCIAL AND OPERATIONAL PRESSURES
Business model and organisation scalability
Competition and health tech market
Intended future use within the healthcare type

Table 3: From [361]: External factors and their dimensions.

according to local standards and practices. These factors affected the transferability of AI-based systems trained on medical datasets. The collaboration at this stage focused on addressing the needs of medical professionals from the intended context of use to support final clinical usefulness across contexts.

Commercial and operational pressures: These are the business and market factors that determine the resources and goals of clinical AI innovation. The available resources affect data collection and the amount of data that can be physically labelled. Meanwhile, market competition may influence choices during the design of the ground truth schemas, e.g., which medical conditions to label.

7.1.3 *Internal Factors*

Two internal factors (Table 4) shape the negotiations between the medical and technical domains over specific labels. They spark debates, discussions, and disagreements that have implications for both the ground truth schema and the resultant datasets.

Epistemic differences: This factor refers to the diverging perspectives, values, and understandings of concepts among domain experts, such as medical professionals, data scientists, and designers. Communication challenges afflicted the close collaboration of practitioners from different domains, which has also been highlighted in Paper I. Moreover, the observed data work was considered part of the "data science" domain. Thus, the organisation and goal-setting remained with the AI engineers. This led to the misapprehension of medical practice and misapprehension of medical knowledge. These misapprehensions manifested in the assumptions about what kind of medical conditions can be labelled and what other information medical professionals can supply. During the collaboration on ground

EPISTEMIC DIFFERENCES
Miscommunication between domains
Misapprehension of medical practice
Misapprehension of medical knowledge
LIMITS OF LABELLING
Domain expert buy-in
Onboarding to the labelling task
Labelling hardware and software
Similarity to the clinical practice

Table 4: From [361]: Internal factors and their dimensions.

truth schemas, medical professionals challenged the assumptions and made changes to the schemas. These changes altered the intended capabilities of the AI model and, in turn, affected the final clinical usefulness of systems trained on this data.

Limits of labelling: This factor refers to the practical and technical constraints that influence the labelling process and the schema. These constraints include domain expert buy-in, onboarding to the labelling task, clinical practice familiarity, and labelling hardware and software. The expert buy-in was a crucial factor that affected what kind of data was collected in the AI4XRAY project. Initially, during the labelling, radiologists marked the location of each finding. This task proved to be too tedious for crucial participants. The task was altered not to collect this metric to keep their contribution to the project. Similarly, the other constraints affected the quality of annotated labels, e.g. the screens used for the labelling task were not diagnostic quality. This means that less experienced labellers may have missed some of the subtle findings. Altogether, these limits altered what kind of data was collected and how it was annotated, affecting the accuracy and consistency of the annotated labels and the schema.

7.2 CLINICAL USEFULNESS: MEDICAL DATASETS CONDITION REAL-WORLD PERFORMANCE, ORGANISATIONAL ACCEPTANCE, AND CLINICAL EFFICACY

This paper foregrounded the extent of factors affecting the creation of medical datasets prior to data labelling. These factors may affect the labels' quality, validity, and reliability, as well as their alignment with medical knowledge and practice. However, their influence does not end on the quality of the collected data. In fact, the factors that shape the datasets indirectly condition the real-world performance, the design space of possible AI functionalities, and their medical focus.

I advocate that the pre-labelling stages of data work, especially the design of ground truth schemas, undergo increased scrutiny by researchers and practitioners. Moreover, it is crucial to increase the

involvement of domain experts and clinical practitioners in the pre-labelling stages of data work. Such collaboration is critical to ensure that AI-based systems trained on the resulting data meet the needs of their intended users by achieving high real-world performance, supporting flexibility during design, e.g., of explainable AI methods, and providing clinically accurate and factual information.

8

SUMMARY OF PAPER III: TOWARDS HUMAN-CENTRED CLINICAL AI

AI innovation has been driven by the implementation of novel models and improving performance captured through technical metrics, sidelining social considerations of use and value delivered in practice [183, 277]. This techno-centric approach resulted in a gap between the needs of end-users and the functionalities offered by contemporary AI-based systems [272, 313, 362]. Consequently, such systems are not widely used in practice [43, 131, 299, 349, 365].

This study was part of the work conducted to support the AI4XRAY innovation project. Aware of the challenges afflicting the realisation of AI in clinical practice [362], I undertook a broad field study to understand radiologists' work and investigate opportunities for AI support across clinical sites and across countries, mindful of the goal of AI4XRAY to make the future system clinically useful in Denmark and Kenya.

I recruited eighteen radiologists and four radiographers from nine medical sites in Denmark and Kenya. Together, I visited two specialised hospitals offering tertiary and quaternary care (DK: 1, KE: 1), five general hospitals providing primary and secondary care and serving as referral centres for specialised treatments (DK: 2, KE: 3), and two imaging clinics, which offered medical imaging and teleradiology services (DK: 1, KE: 1).

During the field study, I conducted observations (DK: 35 hours, KE: 32 hours) and interviews with the participants. The primary goal of the observations was to understand the visited clinic's operations, organisational structure, division of labour, patient demographics and statistics, daily workflows, procedures for handling chest radiographs, and previous encounters with AI-based systems. Moreover, during the observation, when possible, I asked my participants about their visions of useful AI support in regard to the actions they were taking. Finally, I used the interviews to clarify any uncertainties

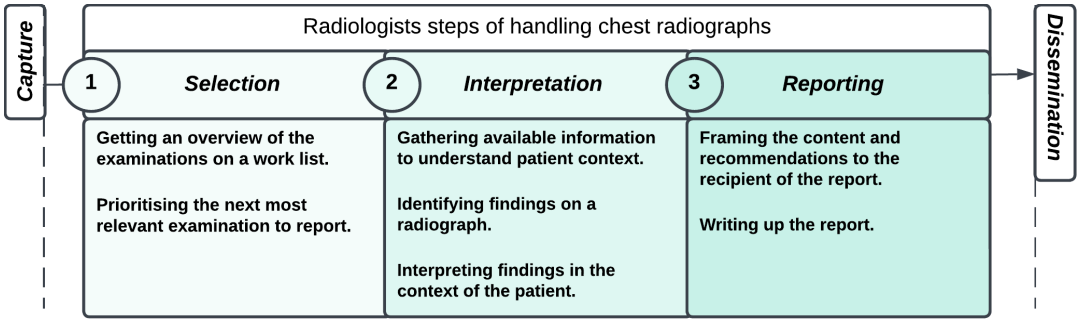


Figure 4: From [360]: A shared workflow of chest X-ray practice observed in Denmark and Kenya.

SELECTION
Challenge: Backlog of X-rays to report exceeding daily processing capacity
Vision: AI distributing examinations by user's expertise
Challenge: Selecting the next most relevant examination to report without an easy overview
Vision: AI detecting medical emergencies
INTERPRETATION
Challenge: Interpreting visually ambiguous findings
Vision: AI providing decision support on subtle and difficult cases
Challenge: Time-consuming process of obtaining additional clinical information
Vision: AI measuring visual features and comparing changes across historical examinations
REPORTING
Challenge: Conveying the right information in a report
Vision: AI double-checking reports against radiographs for missed or misinterpreted findings

Table 5: From [360]: Challenges encountered in chest X-ray practice and envisioned AI support.

pertaining to the areas of observation throughout and after the observations.

8.1 WHAT ARE THE OPPORTUNITIES FOR AI SUPPORT IN CHEST X-RAY PRACTICE?

Similarly to the process in Paper II, to engage in meaningful discussion about the opportunities rooted in practice, I needed first to understand the practice. Surprisingly, the practice of handling chest X-rays was fairly uniform across all the settings in both countries. We observed three main stages: selection, interpretation, and reporting (Figure 4). The differences revealed themselves when comparing work division, type of seen patients, and the expertise of radiologists. However, they were visible along the types of clinics and not countries. This means that the work practices and clinical context were very similar across clinical sites.

Participating radiologists envisioned five ways AI-based systems could realistically help them in practice. These five visions were a response to five concrete challenges (Table 5).

8.1.1 *Opportunities for AI support during the selection of chest radiographs for reporting*

AI distributing examinations by user's expertise: Radiologists who proposed this vision aimed to reduce the backlog of X-rays to report by assigning cases to radiologists based on their level of experience

and specialisation. In the visited sites, the number of chest X-rays to report often exceeded daily capacity. This resulted in delays, stress, and reduced quality of care. Within this vision, examinations deemed normal by AI could be checked by junior radiologists, while examinations deemed by AI to have radiological findings could be reserved for senior radiologists. This way, the AI could optimise the workflow to reduce a backlog. Importantly, this functionality was reserved only for periods of increased workload, not to jeopardise junior radiologists learning.

AI detecting medical emergencies: This vision aimed to support prioritising urgent cases. Radiologists had little overview of the content and context of the X-rays before opening them. They relied on the referrals, urgency attributes provided by ordering clinicians and information about the ordering department. However, these methods usually catch only the most urgent cases and do not account for findings not foreseen by the ordering clinicians. Radiologists envisioned an AI-based system that could flag cases with life-threatening conditions, such as pneumothorax, and move them to the top of their work list. This way, the AI could help radiologists identify critical cases faster and potentially save lives.

8.1.2 *Opportunities for AI support during the interpretation of chest radiographs*

AI providing decision support on subtle and difficult cases: This vision introduces an AI-based system that can provide a second opinion on examinations with visually ambiguous findings or otherwise challenging to interpret. Such challenges are often induced by noise or overlapping structures that make it difficult to identify pathologies. Additionally, cases with subtle or rare findings require more expertise and experience to diagnose correctly. Consultations on such cases are common in clinical sites with several experienced radiologists. However, in smaller settings or in the case of junior doctors who cannot consult their supervisors about every case, such a system could help radiologists resolve doubts, increase confidence, and avoid errors.

AI measuring visual features and comparing changes across historical examinations: Some parts of X-ray examination do not require medical expertise but must be conducted and are time-consuming, e.g., measuring heart-to-chest ratio to exclude enlarged heart diagnosis. Radiologists speculated whether AI could not automatically measure visual features on the radiographs, saving them precious time. Similarly, whenever possible, radiologists compare findings against historical images. In this vision, the AI-based system could also compare the current radiograph with previous ones and highlight any changes or progression of the findings. These functionalities would help radiologists to save time, reduce manual work, and improve accuracy.

8.1.3 *Opportunities for AI support during the reporting of chest radiographs*

AI double-checking reports against radiographs for missed or misinterpreted findings: At this stage, radiologists envisioned a system that could prevent errors and maintain the quality and accuracy of their reports. In practice, radiologists must report all relevant findings to avoid confusion or concern for the clinicians who ordered the X-ray. Additionally, reports need to be tailored to meet the needs and expertise of the recipients, considering factors like their clinical questions and available resources. To support this effort, the envisioned system would analyse a radiologist's report and compare its content against the radiological findings detected in an X-ray. The system would notify the radiologist if any discrepancies or omissions were found. This process would provide an extra layer of quality control without requiring the radiologists to interact with the system actively, thus reducing their workload.

8.2 CLINICAL USEFULNESS: BROADENING AI DESIGN SPACE ENABLES ORGANISATIONAL ACCEPTANCE AND CLINICAL EFFICACY

When evaluating current AI-based systems for radiology, researchers question the actual benefits of using them in practice, pointing to unclear clinical value [206, 313]. This motivated us to start this study by understanding the clinical practice and engaging with radiologists to identify areas where AI could add value. We conceptualised five visions for AI support proposed by our participants and rooted in practice. The unfolded visions contest and expand the traditionally technology-centric outlook on what constitutes valuable AI support. Rather, they centred around practically aiding clinical practice, e.g., by measuring visual features or double-checking reports, thus improving clinical outcomes and supporting radiologists at work.

Overall, envisioning AI support at the point of practice provided new insights about what makes such systems useful. By encouraging end-users to explore their ideas of AI support rather than attempting to retrofit preconceived systems, we may uncover previously unexplored avenues for providing useful support. We advocate that such early grounded engagements can facilitate the transition towards human-centred AI in healthcare.

9

SUMMARY OF PAPER IV: "IT DEPENDS..."

In the Paper III, we discovered new opportunities for AI support in radiological practice. However, successful innovation of the envisioned systems hinges on their final clinical usefulness, i.e., they must deliver robust real-world performance [207, 334, 362], prove clinical efficacy [37, 184, 216], and fit the different intended clinical contexts [62, 67, 169, 360].

Achieving these qualities, and thus clinical usefulness, requires meaningful before-use and in-use configuration of a system to the environment it is going to be used in [135, 169, 263, 356]. The before-use configuration refers to the design decisions taken before a system's integration into clinical practice. In contrast, the in-use configuration refers to the adjustments made after the integration [151].

To explore how configurability may improve the clinical usefulness of AI-based systems in radiology, we conducted nineteen design sessions and interventions with thirteen radiologists across seven clinical sites in Denmark (2) and Kenya (5) (Figure 5). The design sessions were centred around a prototype of an AI-based support system using real data and an AI model developed within the AI4XRAY project. The prototype explored the clinical usefulness of such a system, especially focusing on the aspects that could be configured. Throughout the study, we iteratively improved the prototype based on intermediate feedback. The design sessions were held both online and in person. The design interventions were conducted in-situ to "intervene" in real-world practice. While not using prospective data, we aimed to simulate real use to ground the interaction with the prototype and the feedback received.

9.1 HOW TO CONFIGURE AI-BASED SYSTEMS FOR CLINICAL USEFULNESS ACROSS CLINICAL CONTEXTS?

Based on the data collected during design sessions and interventions, we delineated four technical dimensions of clinical AI that can be configured for clinical usefulness in relation to the social dimensions of the clinical context. The technical dimensions span AI functionality, AI medical focus, AI decision threshold, and AI explainability (Figure 6). The clinical context was defined to comprise five dimensions.

Medical knowledge encompasses the meaning of medical concepts, definitions, and procedures that are relevant to the AI-based system, for example, the meaning of radiological findings. *Clinic type* highlights the differences in patients, resources, focus, and type of offered care between clinical sites of different types. *User expertise level* points attention to the fact that medical professionals have different expertise even if they work at the same clinical site in the same position. *Patient context* comprises their current location (admitted to a hospi-

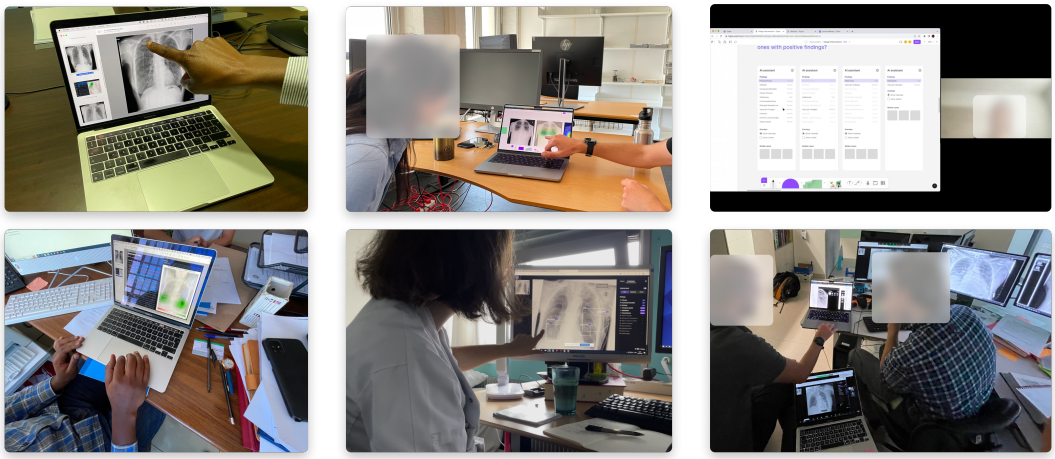


Figure 5: From [363]: Online and collocated design sessions and design interventions with user interface mock-ups and working prototypes (version I, II, II).

tal or not) and medical history. *User situation* describes the workload, available time, and resources of medical professionals using the system.

9.1.1 *Configuration of AI Functionality*

AI functionalities should be aligned with the needs of the local clinic and the expertise of their radiologists. Different clinics may have different workflows, resources, and patient populations, affecting the suitability and applicability of certain AI functionalities. For example, radiologists in a primary care clinic may benefit more from quality assurance than radiologists at a specialised hospital.

AI should be easily accessible whenever necessary or function in the background rather than being an obligatory step in new work practices. End-users should have control over which AI functionalities are integrated into their routine, as clinical work is constantly changing. Moreover, within the same clinical practice, clinical end-users with different expertise levels may benefit from other types of AI support.

9.1.2 *Configuration of AI Medical Focus*

In the context of this study, the AI medical focus refers to the set of radiological findings that the AI detects on chest X-rays.

AI-based systems should be configured to detect the findings typically observed at the local clinical site before use. During this study, we observed that the type of clinic determines the patient demographics seen by radiologists, influencing the prevalence of certain findings.

End-users should be able to select their preferred findings for AI support from the pool of relevant radiological findings. Similar to selecting AI functionality, certain findings may be more challenging for less experienced doctors to assess, while senior radiologists may find support for these findings redundant. Additionally, the clinical significance

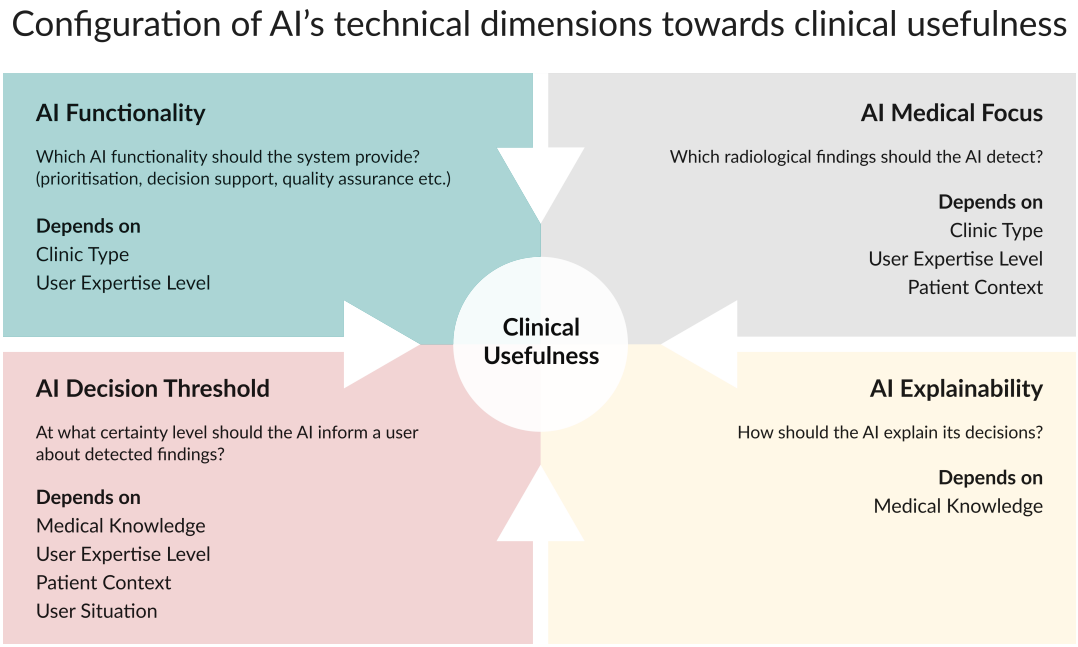


Figure 6: From [363]: A matrix of technical AI dimensions must be configured to achieve clinical usefulness in local practice. The accompanying social dimensions of the local clinical practice condition configuration of each technical dimension. Conceptualised based on design interventions with an AI-based prototype.

of radiological findings varies depending on the patient's context. Integrating this information when defining the medical focus would enhance the fidelity and usefulness of AI support.

9.1.3 *Configuration of AI Decision Threshold*

AI decision threshold is a technical dimension that determines how confident the AI model has to be to make a prediction, effectively influencing the positive and negative predictive values. This value is arbitrarily selected in many AI-based systems to optimise technical performance metrics. We advocate for its more flexible configuration to introduce tailored support and reduce additional work caused by false positive predictions.

Clinical end-users should be able to select a specific threshold per radiological finding. In line with previous configuration needs, certain findings may be perceived as more clinically important. Radiologists could specify the relevancy of specific findings by lowering the threshold, thereby accepting more false positives to ensure the AI system alerts them to these findings. Conversely, raising the threshold value would result in fewer predictions. This functionality could help radiologists manage the work of discerning false positive predictions.

Clinical end-users should be able to adjust the AI decision threshold in relation to clinical parameters (e.g., patient location). From [360], we know that clinical importance depends on factors other than just the definition of a finding. For example, some radiological findings may be more critical if found in an out-patient. To account for that, we sug-

gest that the threshold configuration consider other clinical parameters.

AI decision threshold should be pre-defined to match the requirements of the local clinical site. To further minimise the necessary work by end-users, an AI-based system should be aligned with the type of patients and other requirements of the local clinical site.

9.1.4 *Configuration of AI Explainability*

In this study, we evaluated three distinct methods of explaining AI output on chest X-rays: heat maps, arrows pointing towards radiological findings, and boxes bounding the radiological findings.

Clinical end-users should be able to select their preferred XAI method for each finding, either globally or individually. In general, the methods should allow radiologists to assess the quality of the prediction by inspecting the underlying X-ray and locating the finding. However, the visual representations of radiological findings on chest X-rays vary. This is why a single method for explaining AI predictions in radiology is not a viable solution, e.g., explaining a diffused finding with a fixed box or an arrow may cause consternation and confusion. Rather, selecting the most appropriate method for each finding should be possible.

9.2 CLINICAL USEFULNESS: CONFIGURABLE AI ENHANCES ORGANISATIONAL ACCEPTANCE AND CLINICAL EFFICACY

In Paper III, we explored the link between the work practices of radiologists across different clinical contexts and envisioned opportunities for AI support. However, the described visions are not a panacea that can be applied to boost the clinical usefulness. In Paper IV, we built on the gained knowledge of radiological practice and iteratively developed a prototype to test some of the proposed visions. Particularly, we looked into the practical caveats of configuring an AI-based system to increase clinical usefulness in concrete clinical contexts and scenarios.

We observed that radiologists wished to configure the system to maximise patient benefits and minimise additional workload. In this light, we advocate that clinical usefulness can be enhanced by matching the AI-based systems offering with the characteristics of the local clinical context. Particularly, it is crucial to consider how to make clinical AI systems configurable so that the offered support corresponds to the needs of specific end-users in a specific situation in a local clinical site. We caution that configurability must be considered throughout the innovation process to ensure encompassing configuration before use and appropriate flexibility to configure the system in use.

10 CONCLUSIONS

Surrounded by the hype around AI, I started my PhD in December 2020. During this journey, the hype became louder and louder (looking at you, Large Language Models), and my experiences with AI in clinical practice became bleaker and bleaker. After the ethnographic work and observing how radiologists work across clinical contexts and across countries, I learned virtually no AI-based systems were actually used in practice. I could not help but think, how is it possible that the technology that is said to disrupt virtually every aspect of our lives is barely even used in radiology - one of the most technology-embracing specialities in healthcare?

Within this project, I was given an opportunity to investigate the reasons behind the low use of AI and act on them from the very beginning of an innovation process. I actively participated in shaping the data creation, engaged with real-world medical practices, designed tools and prototypes based on insights from practitioners, and evaluated them in practice. Furthermore, my involvement in the AI4XRAY project enabled me to challenge conventional approaches to AI system development typically seen in the Global North. From the very beginning of the project, I focused my work on crossing boundaries and borders, looking for ways that AI can be beneficial across contexts beyond the confines of a single hospital in Denmark.

The essence of this thesis lies in its unapologetic commitment to innovating an AI-based system driven by clinical usefulness. In Section 5, I presented a comprehensive understanding of clinical usefulness that guided all the other contributions. To recall its definition, it is an overarching quality of AI-based systems emerging from the interplay of their real-world performance, clinical efficacy, and organisational acceptance in a situated clinical context for specific end-users. I also explored how researchers in AI, health, and HCI domains understand and contribute to achieving these constituting qualities. Finally, I contributed to the conceptualisation of clinical usefulness from four new angles.

1. I conceptualised how the expectation of real-world performance depends on the intended use of AI. Particularly how the emphasis on the real-world performance of AI increases, and the value of explainability decreases with the increasing AI agency.

2. I showed how medical datasets condition real-world performance, organisational acceptance, and clinical efficacy. I shed light on the previously invisible pre-labelling work and foregrounded how its considerations condition the final space of AI abilities.

3. I foregrounded how broadening AI design space enables organisational acceptance and clinical efficacy. Specifically, based on grounded envisioning, I conceptualised AI support functionalities that transcend the traditionally evaluated (in radiology) second opinion.

4. I explored how configurable AI enhances organisational acceptance and clinical efficacy. This work showed that even well-thought functionalities need to be configured in relation to social dimensions of clinical work to provide value to end users in practice.

Moreover, I hope this thesis may inform future research and practical innovation projects in clinical AI. Like my studies supported the AI4XRAY project, the answers to the four research questions in this thesis may support the innovation of clinically useful AI-based systems.

RQ1: WHAT ARE THE CHALLENGES OF REALISING AI IN CLINICAL PRACTICE?

I answered this question in Paper I, where we conceptualised five challenges that originate from the dependencies between the social and technical aspects of AI use in clinical practice. The dependencies spanned: training data & ML model and user & system use, system integration & data used and workflow & organisation, user interface and user & system use, user & system use and ML-based system, and healthcare institution & political arenas and use of ML-based system. In this paper, we explored how the challenges of realising AI do not stem from singular issues, like explainability, but are a product of interlinked dependencies. We argued that to address them, there is a need for earlier, closer, and more frequent collaboration between researchers and practitioners from HCI, AI, and health. Additionally, we proposed that early exploration of ideas in clinical practice, e.g., through design interventions, could reveal insights typically acquired after a system has been deployed. This, in turn, could enhance the development of a cohesive vision for the AI-based system, ultimately increasing its chances of being clinically useful.

RQ2: WHAT FACTORS CONDITION THE CREATION OF TRAINING DATA?

In Paper II, we conceptualised five factors that affect the creation of medical datasets. The factors divided between external and internal to the process of labelling spanned regulatory constraints, the context of creation and use, commercial and operational pressures, epistemic differences, and limits of labelling. We deconstructed how these factors affected what data could be collected, the purpose for which it could have been used, and the design of ground truth schemas used to label the data. These very real consequences fundamentally influenced the final shape of medical datasets used for AI training, thus conditioning the capabilities of future AI-based systems.

RQ3: WHAT ARE THE OPPORTUNITIES FOR AI SUPPORT IN CHEST X-RAY PRACTICE?

This question motivated the work described in Paper III. When conducting observations of chest X-ray practice across clinical sites in Denmark and Kenya, we witnessed five problems that afflicted the work of radiologists. These problems prompted our participants to envision five ways AI could help radiologists at work. The visions included distributing examinations by user's expertise, detecting medical emergencies, providing decision support on subtle and difficult cases, measuring visual features and comparing changes across historical examinations, and double-checking reports against radiographs for missed or misinterpreted findings. These visions suggest that the design space of clinically useful AI-based support for radiology has the potential to expand beyond the traditionally seen second opinion. Moreover, their conceptualisation suggests that grounded envisioning may be a relevant method to investigate alternative opportunities for AI support in practice.

RQ4: HOW TO CONFIGURE AI-BASED SYSTEMS FOR CLINICAL USEFULNESS ACROSS CLINICAL CONTEXTS?

Paper IV directly resulted from our work on Paper III. We investigated how to practically realise the conceptualised visions in line with the progress and goals of the AI4XRAY project. To support the usefulness of the visions, we recognised that four key technical dimensions of AI-based systems need to be configured in relation to the five social dimensions of the local clinical context. First, AI functionalities must provide value relevant to the type of intended clinical site and their intended end-users medical expertise level. Second, the findings that an AI-based system for radiology detects should be aligned with the population cared for at the local clinic, end-users medical expertise level, and the context of the patients. Third, the AI decision threshold should respond to the clinical significance of the findings, end users' medical expertise level, current workload, and the context of the patients. Finally, the explainable AI methods should offer visual cues appropriate to the visual representation of explained radiological findings.

10.1 FINAL REMARKS

Together, following my own recommendation from the first paper, I believe that the work described in this thesis embraced the interdisciplinary nature of AI innovation. I hope the description and use of clinical usefulness as the leading quality of AI-based systems will guide future collaborations. Finally, I hope that the four contributions captured in this thesis will generally contribute to the fields of HCI, AI, and health and, more practically, inform the innovation of future clinically useful AI-based systems.

11

WHERE DO WE GO FROM HERE?

EXPLORING THE GENERATIVE ABILITIES OF LARGE LANGUAGE MODELS. It should be noted that, as of February 2024, a notable absence is evident throughout the thesis. In any of the papers, we consider LLMs, which took both the research and popular scenes by storm in late 2022. This, both a choice and a consequence, is related to the AI4XRAY project, which commenced in autumn 2020. At that time, LLMs were but one of the possible research directions and were not considered by the project. Framing the project around imaging data and detection steered the project away from the generative aspects of AI.

I believe that the contributions of this thesis are relevant regardless of the underlying AI technology. However, I also believe that LLMs offer unique abilities that could further enhance AI support. I see investigating how to responsibly incorporate them into clinical practice to empower and provide value to end-users as the next critical step in advancing healthcare innovation.

EXPLORING INTERVENTIONAL METHODS FOR DESIGN OF AI With the ever-faster progress of AI technologies, the conditions for design change as well. New technologies escape traditional design methodologies, as their final capabilities are often realised only in the real context of use. This clashes with the traditional approaches to evaluation in healthcare, resulting in difficulties in obtaining access, risking discouragement of potential end-users or outright rejection of envisioned ideas.

In my thesis, I demonstrated the significance of collaborating with practitioners from the earliest stages of AI innovation. While no prospective design study was conducted, I strove to engage in practice in the real context of use. However, I see this work as a first step towards new interventional methods for AI design. I encourage future research to investigate methods that combine early realisation with the flexibility of design in the context of real-world practice.

Part II

PAPERS

12

PAPER I: CLINCIAN-FACING AI IN THE WILD

TITLE

Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI

AUTHORS

Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, Tariq O. Andersen

DOI

https://doi.org/10.1145/3582430

VENUE

ACM Transactions on Computer-Human Interaction

PUBLISHED

17 March 2023

ABSTRACT

Artificial Intelligence (AI) in medical applications holds great promise. However, the use of Machine Learning-based (ML) systems in clinical practice is still minimal. It is uniquely difficult to introduce clinician-facing ML-based systems in practice, which has been recognised in HCI and related fields. Recent publications have begun to address the sociotechnical challenges of designing, developing, and successfully deploying clinician-facing ML-based systems. We conducted a qualitative systematic review and provided answers to the question: "How can HCI researchers and practitioners contribute to the successful realisation of ML in medical practice?" We reviewed 25 eligible papers that investigated the real-world clinical implications of concrete clinician-facing ML-based systems. The main contributions of this systematic review are: (1) an overview of the technical aspects of ML innovation and their consequences for HCI researchers and practitioners; (2) a description of the different roles that ML-based systems can take in clinical settings; (3) a conceptualisation of the main activities of medical ML innovation processes; (4) identification of five sociotechnical interdependencies that emerge from medical ML innovation; and (5) implications for HCI researchers and practitioners on how to mitigate the sociotechnical challenges of medical ML innovation.

12.1 INTRODUCTION

Artificial Intelligence (AI) is receiving growing attention from policy-makers, scientists, and society at large, as well as massive public and private investments, suggesting a new "AI spring" [219]. The revived interest in AI is also happening in healthcare where the number of publications indexed in PubMed with “AI” in their title has increased almost tenfold in recent years [66]. In this systematic literature review, we focus on Machine Learning (ML) algorithms–a subset of AI that can find patterns and dependencies in complex and unwieldy data [65] without being explicitly programmed [325]. Such algorithms, like deep learning, random forest, and others, have been shown to outperform statistical models when applied to health data (see e.g. [258]). Multiple ML models for the detection and prediction of clinical outcomes have been validated in many disease areas, including diabetes [137], cancer [178], mental health [325], and heart disease [258]. Lab-based studies show that ML may provide promise for clinical practice and clinical outcomes, e.g., predicting health outcomes for improved clinical decision-making [246], reducing diagnostic and therapeutic errors [97, 130, 204, 259], and obviating repetitive clinical tasks [116].

Despite the promising outlook, only few medical devices based on ML have been regulatory approved [238] indicating that clinician-facing ML-based systems are particularly challenging to realise in clinical practice [82, 184]. Going the "last mile" of medical AI innovation is afflicted by many challenges compared to conventional systems [82]. Unlike rule-based algorithms, whose decision-making processes can be inspected, ML models are often criticised of the “black-box” problem [72], i.e., obscuring the reasoning behind a model output [116]. ML requires also large amounts of labelled training data [65, 341], in contrast to conventional statistics-based algorithms. It may also require adjusted modes of interaction, e.g., users participate in improving the algorithm through continuous use [2], or through the use of a system, they take part in the training of a human-in-the-loop ML model. [158].

In Human-Computer Interaction (HCI), Human-AI interaction has been of key interest for more than 20 years [6]. Studies have benefited a user- and human-centred perspective in the pursuit of developing fair, accountable, and useful computer applications that increase automation while augmenting and empowering people [299, 304]. HCI contributions have centred on tackling Human-AI interaction issues by developing conceptual frameworks [336], models and principles [79, 106, 195], methods and techniques [77, 191, 345], and design guidelines [8] for ML-based systems, as well as undertaking experiments with users to empirically explore the challenges of human and AI engagements [169].

Despite HCI’s longstanding contribution to Human-AI interaction, many researchers voice concerns that ML-based systems are uniquely difficult to design, develop, and deploy (see e.g. [355]) in comparison to traditional systems, especially in safety critical and complex settings like healthcare [238, 253, 329]. Challenges arise on various

levels, for example, designers, or HCI researchers and practitioners, find it hard to understand the capabilities of ML [104, 355]; interdisciplinary collaborations between designers, ML engineers, and clinical domain experts require more work than usual [354]; early feedback for iteration is often unavailable in the development phase [77, 356]; and the unpredictable outcomes of ML models or "imperfect algorithms" make it difficult to obtain purposeful and trustful interfaces for prototyping ML [191, 355]. Many issues arise because of the dynamic nature of ML models and the challenges of interdisciplinary collaboration in the innovation process of designing, developing, and deploying ML-based systems. These issues have been attributed to inherently different methodological approaches among researchers and practitioners from HCI, ML, and Health [43, 132, 343, 355] and the lack of mutual understanding, shared conceptual frameworks, and well-established interdisciplinary methods and techniques [43, 132, 343, 355].

More recently, scholars have engaged with designing, developing, and deploying ML-based systems for healthcare settings called for revisiting and re-framing the problem of designing *Human-AI interaction* as a matter of designing an ML-based *sociotechnical system* [26, 82, 109, 169, 221, 253, 299, 325]. In their systematic review of HCI literature on ML in mental health [325], the authors caution that the development of effective and implementable ML systems *"is bound up with an array of complex, interwoven sociotechnical challenges"*. Research shows how integrating an ML-based system in a clinical setting may result in breakages of social structures that require "repair work" [26], and how an ML model with high accuracy in the lab was heavily impacted in practice by "socio-environmental" factors like clinical workflows and patient experience [26]. These and several other studies form an important emerging discourse in HCI that warns against developing ML models apart from clinician-facing systems that incorporate them and without the involvement of end-users. These systems should rather be understood as "complex sociotechnical system[s]", which only become effective through an innovation process that successfully inter-works *"complex sets of people, practices, technologies, and infrastructures"* [109].

In this review, we follow the sociotechnical turn and acknowledge the lack of well-established conceptual frameworks, models, principles, methods, and techniques supporting the interdisciplinary design, development, and deployment processes of clinician-facing ML-based systems. Our analysis builds on literature that qualitatively explores real-world implications of clinician-facing ML-based systems and engages with the issues encountered during medical ML innovation. Throughout the paper, we use *innovation* to describe all the activities from the conception of an idea for an ML-based medical system to its use "in the wild" [284]. We posed 6 research questions (RQ) to guide this review and form a basis for our contributions: a conceptual framework (1-4) to support medical ML innovation, and opportunities (5) for how to tackle challenges faced during the realisation of clinician-facing ML-based systems.

1. Overview of the technical aspects and their consequences for HCI researchers and practitioners
   RQ.1 What kind of health data, ML algorithms, integration methods, and ML development approaches are present across the literature?

2. Description of the different roles that ML-based systems can take in clinical settings
   RQ.2 What are the intended uses of ML-based systems in clinical settings?

3. Conceptualisation of the main activities of the medical ML innovation process
   RQ.3 Which main activities were reported, and what were their purposes in the ML innovation processes?
   RQ.4 Who were the end-users and other stakeholders, and how were they involved in the ML innovation processes?

4. Identification of five key sociotechnical interdependencies that emerge from medical ML innovation.
   RQ.5 Which sociotechnical challenges were encountered during the use of medical professional-facing ML-based systems?

5. Implications and opportunities for HCI researchers and practitioners on how to navigate and contribute to the complex collaborative space for innovating such novel systems.
   RQ.6 How can HCI contribute to the medical ML innovation processes

In the following section, we describe the systematic review method. The results section includes four main sections where we analyse data extracted from the included articles to address the research questions. We follow the results by contextualising them within the broader HCI literature. Lastly, we highlight five opportunities for HCI researchers and practitioners: how to engage in interdisciplinary collaboration during the ML innovation process; and how to tackle the sociotechnical challenges that afflict it.

12.2 MATERIALS AND METHODS

12.2.1 *Search strategy*

The literature search was completed on April 7, 2021. We searched three databases: ACM DL, PubMed, and arXiv. Given the translational nature of our focus, we included outlets aggregating studies from Computer Science, Health, and unpublished work. We included all types of publications written in English and did not constrain the publication date. To include the widest possible array of studies, we constructed queries for each of the databases in collaboration with an information specialist from the Royal Danish Library. Respective queries are attached as appendices A.1.1 – ACM DL, A.1.2 - PubMed, and A.1.3 – arXiv. The queries returned 9,672 publications eligible for screening (ACM = 4109, PubMed = 5561, arXiv = 37). We included

one external publication that was not returned by the queries but was known to the authors of this paper. It was submitted to JMIR in January 2021 and accepted in October 2021.

12.2.2 *Selection process*

The selection process took place between April 7, 2021, and September 23, 2021. Five authors took part in the process (HDZ, DL, XD, FK, TOA). We used Rayyan.ai [254] to conduct the screening process, which comprised 3 phases. A flow diagram of Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [255] of the process can be seen in Figure 7. The selected articles had to meet six inclusion criteria described in Table 6. Ultimately, 25 publications were selected for the review.

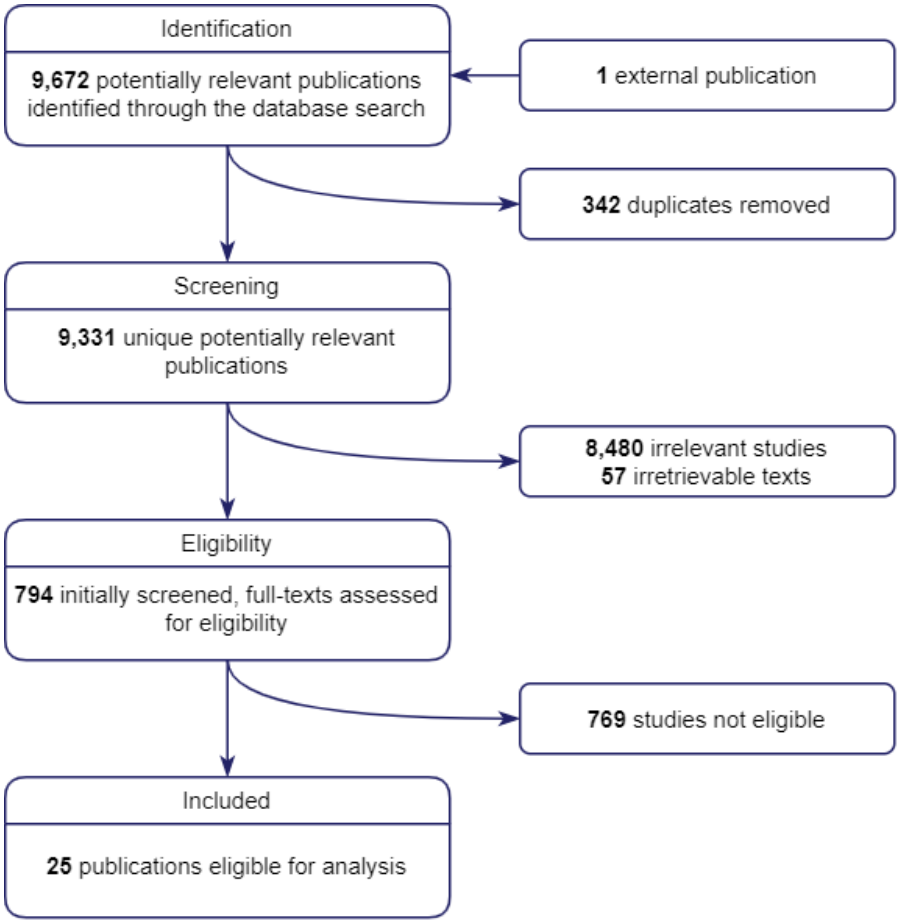


Figure 7: A PRISMA flow diagram of the literature search and study selection

Table 6: Eligibility criteria.

CRITERION	INCLUSION	EXCLUSION
Machine Learning-based	Studies that described systems that used machine learning. We considered machine learning as algorithms that make predictions or decisions without being explicitly programmed [194, 325]. Relevant algorithms included but were not limited to neural networks, random forests, genetic algorithms, Bayesian networks, support vector machines.	Studies that described rule-based systems, knowledge bases, or conventional statistical models that rely on domain experts to *"specify the variables that are relevant for a particular analysis."* [325]
ML produces relevant medical outcomes	Studies that described medical systems affecting patients' clinical outcomes e.g. pathologies detection, treatment suggestion, and health state prediction.	Studies that described medical systems that were not directly related to patients' clinical outcomes e.g. cost estimation, designing and following medical guidelines, administrative task automation, International Classification of Diseases (ICD) codes prediction, or model comparisons.
Investigates implications for clinical practice	Studies that reported on implications of ML-based systems for clinical practice, e.g., evaluation of a system in use [334], perspectives on interaction design in the context of clinical use [67], and investigation of barriers to successful clinical use [261].	Studies that focused on an ML-based system outside of the scope of clinical use, e.g. solely in a patient's home [94].
Includes some version of the system	Studies that included a system, prototype, or a mock-up thereof. Any data created in relation to ML should stem from interactions with a concrete system.	Studies that described theoretical and non-realised systems.
Collects subjective and qualitative data	Studies that reported on information about sociotechnical aspects of ML implementation in healthcare.	Studies that reported solely on technical metrics.
Involves medical professionals	Studies that involved healthcare professionals.	Studies that included only patients, administrative staff, IT specialists, or dentists.

12.2.3 *Data extraction and synthesis*

In our analysis, we strove to follow qualitative systematic review principles [58, 99]. Our first working principle was iteration. Due to the nature of the sought-after insight, we were unable to extract all the data using pre-piloted forms. Instead, HDZ and TOA conducted a

preliminary analysis of several relevant articles using the constructing grounded theory approach [74] in NVivo 13 (QSR International). The two authors coded line-by-line excerpts relevant to the inclusion criteria and the research questions. No code book was present and ambiguous excerpts were coded. Interpretative categories resulted from axial coding. Not all of the initial codes were assigned to a conceptually richer category. At the end of the preliminary analysis, we created an open codebook that served as the starting point of the main analysis and the initial research questions.

During the main analysis, we followed the procedure from the preliminary stage and used the previously developed codebook to increase code fidelity. The open coding was followed by an axial coding phase, which resulted in several levels of grounded concepts. We did not limit ourselves to the pre-existing categories but rather used them to search for previously unmentioned concepts. Throughout the review process and especially during synthesis work, we repeatedly returned to the original texts to extend and modify the codebase. We sought for the first-level codes to be grounded in the data instead of our interpretation. This was especially important towards the end of the analysis, as our understanding of the described phenomenon was deeper, and we were abstracting the data. Hence, we continuously compared the codes against the corresponding quotes to ensure their verity.

Based on the first-level codes, we produced several high-level synthetic categories like "data and feature quality" and "accuracy and performance". We followed an iterative process of coding, refinement, reorganisation, and redefinition. For example, when analysing sociotechnical challenges, we moved between creating synthetic categories and grouping into sub-insights, followed by adding an interpretive layer to the themes that emerged, e.g., "AI algorithms and data" became a relation between "Training Data & ML Model and User & System Use." During the overall analysis process, we continuously revisited the main research questions as new insights emerged. We did this to ensure a connection between the data and the findings that we deemed relevant and interesting for HCI. Meetings between three authors (HDZ, TOA, FK) to discuss the synthetic categories, their underlying codes, and the relationships between them formed the backbone of the process. In our discussion, we took a reflexive stance to remain critical of our interpretation and to stay true to the original authors' meaning.

12.3 RESULTS

12.3.1 *Studies Overview*

Included articles targeted predominantly Health outlets. Out of 25 included publications, 13 were published in Health Informatics, 4 in Health, 6 in Human-Computer Interaction, and 2 in Machine Learning outlets. A majority of the articles were published in 2020 (11),

followed by 2019 (6), 2021 (3), 2017 (2), 2018 (1), 2014 (1), and 2004 (1). The total distribution can be seen in Figure 8.

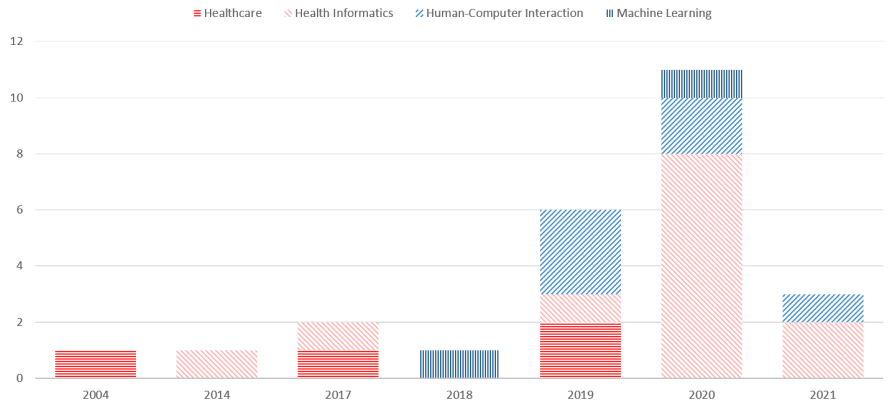


Figure 8: A distribution of the included articles grouped by publish year and highlighted domains

The 25 included articles describe 23 distinct clinician-facing ML-based systems, as shown in Table 7. The only system described in more than one publication that met our criteria was Sepsis Watch [290, 301, 302]. Although medical specialities in focus were not always explicitly stated in the article text, we derived them based on the authors' description of (1) the place of the study - clinical facility or department; (2) the focal condition; and (3) the involved medical professionals. We did not find a medical speciality that was more receptive to ML innovation than others. The speciality targeted by most systems (3) was emergency medicine, and 5 systems were classified as targeting interdisciplinary settings.

Table 7: Overview of included studies. Authors' evaluation denotes original authors' high-level assessment of the overall results of a study. Adoption refers to systems in clinical use and comprises two types: low and high, which denote the degree to which the majority of the end-users used it in their daily practice. Response refers to systems at a stage before clinical use and comprises two types: positive and mixed, which reflect users' sentiments towards the systems.

AUTHOR	MEDICAL SPECIALITY	INTENDED USE	ML OUTPUT	SYSTEM STATUS	AUTHORS' EVALUATION
Barda et al. [22]	Paediatrics	Prioritisation Decision support	Prediction (Overall mortality risk)	Pre-deployment Prototype	Positive responses
Baxter et al. [25]	Internal medicine	Prioritisation	Prediction (Overall unplanned readmission risk)	In clinical use Operational system	Ineffective adoption
Beede et al. [26]	Ophthalmology	Automation (Issuing referrals)	Detection and severity rating (diabetic retinopathy)	Deployment Operational system	Ineffective adoption
Benda et al. [28]	Interdisciplinary	Prioritisation	Prediction (potential cost and care needs)	Pre-deployment Mock-up	N/A

Table 7

AUTHOR	MEDICAL SPECIALITY	INTENDED USE	ML OUTPUT	SYSTEM STATUS	AUTHORS' EVALUATION
Brennan et al. [57]	Surgery	Decision support (Risk assessment)	Prediction (Eight types of postoperative complications)	Deployment Prototype	Mixed responses
Cai et al. [68]	Laboratory medicine Oncology	Decision support (Diagnosis)	Detection and severity rating (Prostate cancer)	Pre-deployment Prototype	N/A
Cai et al. [67]	Laboratory medicine Oncology	Decision support (Case-based reasoning)	Similar historical cases (Prostate cancer)	Pre-deployment Prototype	Positive responses
Cho et al. [80]	Clinical care	Prioritisation Decision support (Prevention of adverse effects)	Prediction (Falling risk) Intervention recommendations	In clinical use Operational system	Mixed adoption
Gastounioti et al. [122]	Vascular medicine	Decision support (Diagnosis)	Prediction (Atherosclerosis risk) Similar historical cases	Deployment Operational system	N/A
Ginestra et al. [127]	Emergency medicine	Prioritisation	Prediction (Sepsis risk)	In clinical use Operational system	Ineffective adoption
Gu et al. [134]	Oncology	Decision support (Case-based reasoning)	Similar historical cases (Breast cancer)	In clinical use Operational system	Positive responses
Hollander et al. [155]	Cardiology	Decision support (Diagnosis)	Prediction (Acute coronary syndrome risk and acute myocardial infraction)	In clinical use Operational system	Ineffective adoption
Jauk et al. [171]	Interdisciplinary	Prioritisation	Prediction (Delirium risk)	In clinical use Operational system	Ineffective adoption
Jin et al. [172]	Interdisciplinary	Decision support (Diagnosis, treatment, case-based reasoning)	Diagnosis suggestion Intervention suggestions Treatment simulation Similar historical cases	Pre-deployment Operational system	Positive responses
Matthiesen et al. [221]	Cardiology	Prioritisation Decision support	Prediction (Ventricular tachycardia and ventricular fibrillation)	Pre-deployment Mock-up	Mixed responses
McCoy et al. [223]	Emergency medicine	Prioritisation	Prediction (Sepsis risk)	Deployment Operational system	Successful adoption
Morrison et al. [237]	Neurology	Decision support (Diagnosis)	Detection and severity rating (Multiple Sclerosis)	Pre-deployment Prototype	Positive responses
Petitgand et al. [261]	Emergency medicine	Decision support	Compilation of relevant medical information	Deployment Operational system	Ineffective adoption

Table 7

AUTHOR	MEDICAL SPECIALITY	INTENDED USE	ML OUTPUT	SYSTEM STATUS	AUTHORS' EVALUATION
Romero-Brufau et al. [286]	Interdisciplinary	Prioritisation	Prediction (Overall unplanned readmission risk)	Deployment Operational system	Successful adoption
Romero-Brufau et al. [285]	Internal medicine	Prioritisation Decision support	Prediction (Diabetes risk) Intervention recommendations	In clinical use Operational system	Ineffective adoption
Sandhu et al. [290]	Emergency medicine	Prioritisation Decision support	Prediction (Sepsis risk)	In clinical use Operational system	Successful adoption
Sendak et al. [301]	Emergency medicine	Prioritisation Decision support	Prediction (Sepsis risk)	Deployment Operational system	Successful adoption
Sendak et al. [302]	Emergency medicine	Prioritisation Decision support	Prediction (Sepsis risk)	Deployment Operational system	Successful adoption
Wang et al. [334]	General practice	Decision support (Diagnosis, treatment, case-based reasoning	Diagnosis suggestions Intervention recommendations Similar historical cases Knowledge base	In clinical use Operational system	Ineffective adoption
Yang et al. [356]	Cardiology Surgery	Decision support	Prediction (Postoperative lifespan)	Pre-deployment Mock-up	Mixed responses

Between the different Health domains, the ML-based systems delivered different types of outputs. 14 of the projects evaluated ML that predicted a diverse set of medical events relevant to the medical speciality in focus. 4 studies focused on the use of ML to retrieve similar historical cases, and 3 projects involved ML that detected and rated the severity of a focal condition. Notably, the same type of ML output was used to serve different purposes and thereby different *intended uses* of ML.

Intended use can be understood as the primary purpose of an ML-based system and the reasons for using ML to reach that goal. Below, in section 3.3 we provide an in-depth conceptualisation of *intended uses*. Across the articles, we identified 3 main *intended uses* of ML in a medical setting: (1) decision support, (2) prioritisation, and (3) automation of tasks. Many of the systems supported more than one *intended use*. 18 systems focused on decision support, 9 systems utilised ML for prioritisation, and a single one was used for automation. This distribution suggests that using ML for decision support in medical settings is more mature and better understood than using ML for prioritisation and automation of clinical tasks.

To provide an overview of the high-level results of the studies, we reported on systems' adoption or responses from clinical end-users. Out of the 23 distinct systems, 12 were deployed and evaluated in the wild. While only 3 deployments were deemed successful by the authors, some of the remaining systems may see increased use in the future, e.g. Jauk et al. [171] outlined several steps to increase

adoption. Pre-deployment evaluations received better responses on average. No system received solely negative responses. 5 out of 10 systems received positive responses, and two lacked an overall assessment. We opted for the high-level results due to the high variability of the reported metrics and their complex influence on adoption and responses. The metrics reported in the articles included, among others: F-score [134], time spent on a task [122], sensitivity and specificity [134, 286], area under the receiver operating characteristic curve [57, 134], and Positive Predictive Value (PPV) [286], i.e., the probability that disease prediction corresponds to a patient having that disease.

12.3.1.1 *Intended use affects performance needs*

Articles that described systems primarily used for *decision support* tended to pay greater attention to the explainability of their output rather than perfect predictive values. Providing contributing factors, raw input data, and historical context helped clinicians to *"explore and interpret prediction results for better clinical decisions"* [172] and augmented their current clinical skills while providing previously unavailable resources [67, 237]. In some cases, the lack of such additional information reduced trust in the predictions [334, 356]. Similarly, Morrison et al. [237] pointed out that simply displaying the algorithmic decision-making process may lead to misunderstandings and confusion among the end-users due to radically different reasoning between humans and algorithms.

When a system was used primarily for *prioritisation*, researchers paid great attention to positive predictive value, even at the cost of explainability [301, 302] or other technical metrics, e.g., with high PPV, even a relatively low sensitivity (67%) was deemed satisfactory [286]. Baxter et al. [25] argued that high PPV is pivotal, as false positives require additional work and resources. A similar point was raised by clinicians interviewed by Matthiesen et al. [221] who reported terminating the clinical use of an alert in a remote monitoring system, which had a high false positive rate. While explainability was often not the primary quality sought after when *prioritising*, presentation of additional information, e.g., raw input data, or factors that had a significant positive and negative effect on the prediction, was considered beneficial to the end-users [301, 302].

Lastly, *automation* evinced the highest reliance on predictive performance among the three intended uses. In the single study that focused on this intended use, low PPV was tied to an increased need for resources which cost had to be borne primarily by the patients [26]. However, in contrast to *prioritisation* and *decision support*, the cost of low negative predictive value could have been even higher. After the pilot, would the system overlook a case of retinopathy, the patient would miss a potentially sight-saving referral.

12.3.2 *Technical Overview*

Machine Learning has a profound influence on a system's success in the wild. Different types of ML algorithms offer varying capabilities and impose varying limitations. They not only produce different types of outputs but also reach their conclusions in different ways. Lack of technical understanding by HCI practitioners is one of the known challenges in the meaningful conceptualisation of ML-based systems [355]. Based on the reviewed articles, we distinguished four technical aspects that require varying resources and distinct consideration from HCI practitioners during the innovation process: type of the ML algorithm used, type of the data used, integration method, and *ML development approach*. The overview can be seen in Table 8.

Table 8: Overview of machine learning algorithms, used data, and ML development approach.

Data	Structured (EHR data)	Barda et al. [22] Baxter et al. [25], Cho et al. [80], Gastounioti et al. [122], Ginestra et al. [127], Gu et al. [134], Hollander et al. [155], Jauk et al. [171], Jin et al. [172], McCoy et al. [223], Romereo-Brufau et al. [285], Romero-Brufau et al. [286], Sandhu et al. [290], Sendak et al. [301], Sendak et al. [302], Wang et al. [334], Yang et al. [356]
	Image (retina image, ultrasound image, mammogram, physical examination recording, biopsy image)	Beede et al. [26], Cai et al. [67], Cai et al. [68], Gastounioti et al. [122], Gu et al. [134], Morrison et al. [237]
	Text (sociodemographic data, insurance claims, patient's self reported data)	Romero-Brufau et al. [285], Romero-Brufau et al. [286], Benda et al. [28], Petitgand et al. [261]
	Time series (defibrillator implants transmissions)	Matthiesen et al. [221]
Machine learning	Knowledge-driven (random forest, SVM)	Barda et al. [22], Ginestra et al. [127], Jauk et al. [171], Matthiesen et al. [221], Morrison et al. [237], Romero-Brufau et al. [285], Gastounioti et al. [122]
	Data-driven (deep learning, neural network)	Beede et al. [26], Cai et al. [68], Cai et al. [67], Jin et al. [172], Petitgand et al. [261], Sandhu et al. [290], Sendak et al. [301], Sendak et al. [302], Hollander et al. [155]
	Implementation dependent (bayesian network)	Cho et al. [80]
	Other (genetic algorithm, custom classifier, n/a)	Gu et al. [134], Baxter et al. [25], Brennan et al. [57], Romero-Brufau et al. [286], Wang et al. [334], Yang et al. [356], Benda et al. [28], McCoy et al. [223]

Table 8

Integration method	Standalone application (Web application, PC program)	Beede et al. [26], Brennan et al. [57], Cai et al. [68], Gaastounioti et al. [122], Jin et al. [172], Morrison et al. [237], Romero-Brufau et al. [285], Romero-Brufau et al. [286], Sandhu et al. [290], Sendak et al. [301], Sendak et al. [302]
	EHR	Barda et al. [22], Baxter et al. [25], Cho et al. [80], Ginestra et al. [127], Jauk et al. [171], Wang et al. [334]
	Other (Printouts, phone call)	Hollander et al. [155], Matthiesen et al. [221], McCoy et al. [223], Petitgand et al. [261], Yang et al. [356]
	N/A	Benda et al. [28], Cai et al. [67], Gu et al. [134]
ML development approach	Cohesive	Barda et al. [22], Ginestra et al. [127], Gu et al. [134], Jin et al. [172], Matthiesen et al. [221], Sandhu et al. [290], Sendak et al. [301], Sendak et al. [302]
	Discrete	Beede et al. [26], Benda et al. [28], Brennan et al. [57], Cai et al. [68], Hollander et al. [155], Jauk et al. [171], Morrison et al. [237]
	Third-party	Baxter et al. [25], McCoy et al. [223], Romero-Brufau et al. [285], Romero-Brufau et al. [286]
	N/A	Cai et al. [67], Cho et al. [80], Petitgand et al. [261], Wang et al. [334], Yang et al. [356]

12.3.2.1 *Types of medical data*

Most popular machine learning algorithms are developed for specific data types, e.g., convolutional neural networks have become dominant in various computer vision tasks, whereas recurrent neural networks mainly are used for modelling sequential data, such as text. Different data types pose various challenges to innovation processes. We categorised four data types: structured, image, text, and time series. Each data type has different qualities relevant that influence the choice of an ML algorithm, design opportunities, and in-the-wild use.

The majority of the described systems (15) used structured data accessible through an Electronic Health Record (EHR) system. While this data type is characterised by semantic richness and objectivity, it often offers limited datapoints per patient. For example, Polubriaginof et al. [266] assessed the quality of family history data captured in an established commercial EHR at a medical centre. After analysing differences between 10,000 free-text and 9,121 structured family history observations, they found that free-text notes contain more information than structured ones. Moreover, the variance of datapoints between similar patients may also vary, e.g., *"quality of the EHRs collected in Chinese hospitals were much worse than those of the MIMIC dataset"* [172]. A final challenge when using structured EHR data is availability. In comparison to other data types like image or text, available

training data is sparse, which hinders progress in developing the ML algorithms that rely on it.

6 systems utilised image data ranging from single retina images [26] to videos from a depth camera [237]. This data type is characterised by objectivity, and format consistency between patients, however, the quality of the input may vary [26]. Thanks to the wide availability of training data, ML algorithms that use image data are relatively advanced and offer high performance.

5 systems relied on text data as their input. Similar to image-based algorithms, these algorithms benefit from a plethora of training data, which results in many high-performing models available to the researchers. However, this type is also burdened with nuance and noise that must be examined. Clinical notes are usually written by practitioners under time pressure, they include a lot of abbreviations and jargon that result in challenging transferability [145]. In the reviewed articles, text data was accessed through public and private databases of sociodemographic data based on a patient's postal code or Medicare claims (in the USA) [28, 285, 286]. A common denominator of these studies was the heavy influence of social determinants on their focal conditions. A different approach was employed by Petitgand et al. [261], who analysed patients' self-reported data.

Lastly, a single article reported on using time series data, which, like text data, can be described as a noisy sequence. Unlike text, time series data suffers from limited available training data: *"11,921 transmissions from 1,251 patients ... followed over a 4-year period."* The low number of transmissions resulted in the need for collaborative feature engineering with domain experts and low performance of data-driven algorithms [221].

12.3.2.2 *Machine learning algorithms*

We divided the reported ML-based systems based on their dependency on domain (i.e., medical) expertise: knowledge-driven, data-driven, implementation-dependent, and others. Such division not only exemplifies the varying needs — including data needs — during the innovation process, but also, e.g., informs about their explainability potential.

We identified 7 out of 23 ML algorithms as knowledge-driven because they show more direct dependency on domain knowledge elicited from experts, e.g., through feature engineering [121, 350]. Engineering informative features for available resources signifies that these algorithms do not have to derive them implicitly from the data. Thus, knowledge-driven ML algorithms require, in general, smaller labelled datasets to produce accurate output. Moreover, knowledge-driven algorithms provide substantial opportunity for explainability, as weights associated with manually designed features can be used to explain their output [309], e.g. *"The dataset ... consisted of 11,921 transmissions. The Random Forest ML method was selected ... because it provided optimal results when considering the tradeoffs between*

model performance and explainability ... show top five features that increase or decrease the likelihood of [an ...] event" [221].

7 out of 23 systems used a data-driven ML algorithm. This type of algorithm does not require direct knowledge input from domain experts to engineer data features. Instead, it derives them from annotations provided by domain experts [203]. Due to this, they require much larger data sets compared to knowledge-driven algorithms. Because they derive features automatically, they may not be easily understandable to humans; hence, their output explainability is not straightforward and remains an active research area in the machine learning community [308]. However, in cases where domain experts cannot be involved to a satisfactory degree or it is difficult to analyse multi-dimensional data, data-driven algorithms offer better performance. As remarked by Sendak et al. [301] *"dataset contained over 32 million data points ... a broad range of features, including medical history and all repeated vital sign and laboratory measurements ... model explainability was not prioritized ... The model extended prior work using recurrent neural networks"*.

9 out of 23 articles described ML algorithms that could not have been assigned to any of the above categories, either due to the versatility of the algorithm (bayesian network [80]) or the lack of technical description beyond claims about using machine learning.

12.3.2.3 *Integration method*

The third technical aspect relevant to the design, development, and in-the-wild use is access to an ML output. In this review, we call it an integration method. Its relevance was described by one practitioner in the following words: *"[m]ake sure that's in their workflow. If you expect someone to go to a third-party system or to a website, you've lost"* [28]. The described systems were integrated into clinical workflows in three distinct modes: as a standalone application, within a hospital's EHR system, or in other (oftentimes analogue) ways.

Although 15 systems relied on EHR data as their data source, only 6 of them presented their output within corresponding EHR instances. In fact, standalone systems prevailed as the preferred workflow integration method. 9 out of 23 systems were accessible via a web application or a PC programme. The rest of the systems opted for alternative methods, e.g., Petitgand et al. [261] and Yang et al. [356] used printouts as a way to communicate the ML algorithm's predictions. A single system's output was communicated via a phone call to the attending physician [223]. In a few instances, the authors did not report in detail on the integration method.

12.3.2.4 *ML development approach*

ML development approach is a term we introduced to denominate the relationship between the development of an ML algorithm and the software embedding it. We distinguished two approaches – cohesive and discrete. A cohesive approach informs us that the development of

an ML algorithm is an integral part of the overall ML innovation process and that it is subject to change along with the final system. Conversely, in a discrete approach an ML algorithm is developed prior to and independently of the final system. In this approach, the ML algorithm becomes a virtually immutable part of the final system.

Based on the accounts of various development activities, we concluded that the cohesive approach was applied 7 times, the same number of times as the discrete approach. Out of all the described systems, 9 of them cannot be assigned to any of the categories. In 5 cases, the original authors did not supply enough information to determine the approach, or the study took place before development. Similarly, we did not assume the *ML development approach* of systems supplied by third-party vendors.

12.3.3 *Intended Uses of Clinician-Facing ML-Based Systems*

We distinguished three overall roles of clinician-facing ML-based systems in clinical practice: decision support, prioritisation, and automation. We employed the concept of *intended use* to scrutinise the diversity of these roles. It has also been used in the engineering domain of medical device development, where it is an important part of receiving regulatory approval [238]. However, in the following, we used it to encapsulate the sociotechnical intricacies of clinicians' needs, systems requirements, clinical utility, and its situated use. This way, the *intended use* connects the technical choices described above (i.e. *ML development approach*, machine learning algorithm, and integration method) with the intended purpose of the ML-based system and its use in clinical practice.

12.3.3.1 *Decision support*

A clinician-facing ML-based system designed for decision support assisted healthcare professionals in decision-making within the context of a single patient (Figure 9). The contributions of ML-based systems during that process can vary. We recognised two main types of assistance reported in the articles, which were not mutually exclusive and were used in concord: alternative outlook and quality assurance.

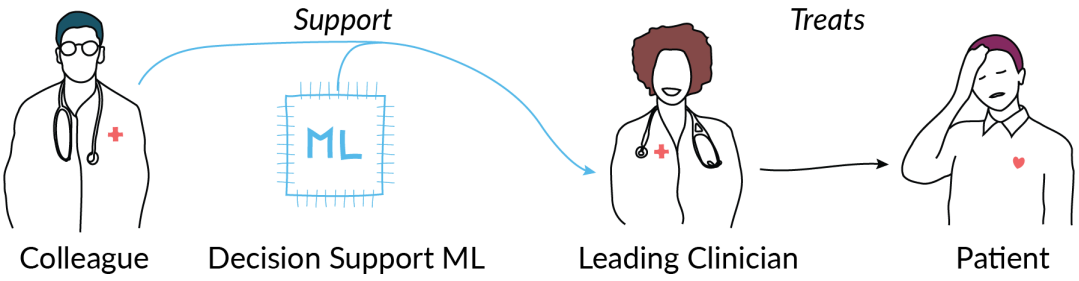


Figure 9: A clinician-facing ML-based system supports clinicians at work.

Alternative outlook denotes use cases where ML adds to the understanding of a patient's condition. We distinguished five different subtypes that stem from the reviewed articles. *Quality assurance* includes any *intended use* where ML acts as a safety net helping providers avoid mistakes. We distinguished three different sub-types grounded in the reviewed articles. The types are described in detail in Table 9.

Table 9: Descriptions and examples of the intended use: decision support.

DECISION SUPPORT: ALTERNATIVE OUTLOOK
ML diagnoses patients ML helps providers make a correct diagnosis. Cai et al. [67] reported on a system that they could consider a peer who provides a second opinion about the presence and severity of prostate cancer. Similar systems, though with less agency, were described by Gastounioti et al. [122], Hollander et al. [155], and Morrison et al. [237]. In their scenarios, physicians received the probability of specific diagnoses the systems were designed to detect. Jin et al. [172] and Wang et al. [334] reported on more advanced systems that predicted the probability of multiple conditions. That ability was remarked by one of the evaluating physicians, as especially useful to novice doctors [172].
ML finds similar historical cases The reported accounts of ML-based systems that retrieved similar patients contributed two-fold to the diagnostic process. First, historical data provided an outlook on similar patients' development, alternative treatment methods, or diagnoses [68, 134, 172]. Second, Cai et al. [67] reported that physicians reflected on their judgement when the system presented patients they did not expect to see.
ML presents available information in a new way Morrison et al. [237] described how offering a new perspective on existing data led to a better diagnosis. The authors achieved that by comparing focal patient data to historical data points of a larger cohort.
ML provides new information Brennan et al. [57], Jin et al. [172], Yang et al. [356] explored ML-based medical systems that generated a completely new type of data. Based on historical data, ML returned a prediction of the future state of a patient. The prediction was bound to a focal medical intervention [57, 356] or could be triggered by selecting an intervention to see its predicted impact on a patient [172].
ML provides intervention recommendations Among the reviewed applications, we discovered two types of recommendations: preventive interventions and treatments. Benda et al. [28], Cho and Jin [80], and Romero-Brufau et al. [285] described systems that estimated risks of certain adverse events and helped mitigate those risks. Responses collected by Benda and colleagues conveyed that providers accepted suggestions when the adverse event was a complex issue [28]. Similarly, Jin et al. [172] and Wang et al. [334] evaluated systems that suggested treatments to described conditions.

Table 9

DECISION SUPPORT: QUALITY ASSURANCE
ML double-checks provider's decision Such systems can assess providers' decisions or diagnoses, e.g., for adverse effects. This use-case was reported in two studies [172, 334]. Similar to asking a peer for a second opinion on a diagnosis, physicians consulted the ML to check for unanticipated adverse effects or other mistakes.
ML provides consistent diagnoses In certain conditions, a definitive diagnosis was not possible, e.g., due to the lack of unequivocal testing methods (sepsis [127]) or a high degree of subjectivity in the existing ones (multiple sclerosis [237]). In such situations, ML-generated diagnoses provided consistent output that served as a reference point for providers making a decision.
ML enhances data access Healthcare providers had often limited time to make a diagnosis [172]. Two articles reported on clinician-facing ML-based systems that expedited access to relevant health information. This was achieved by refining condition qualities, which eased the identification of similar patients [67], and pre-screening of patients to highlight historical information relevant to the current condition [261].

12.3.3.2 *Prioritisation*

An ML-based system designed for prioritisation assists healthcare providers within the context of multiple patients (Figure 10). It conducts an individual assessment of all the relevant patients according to predefined criteria. The outcome of such assessment is available to medical professionals and helps them plan their work. The goal of the ML-based prioritisation system is to alter the order of providers' actions and highlight who needs their attention. We distinguished four sub-categories of the types of additional information supplied by the systems (Table 10).

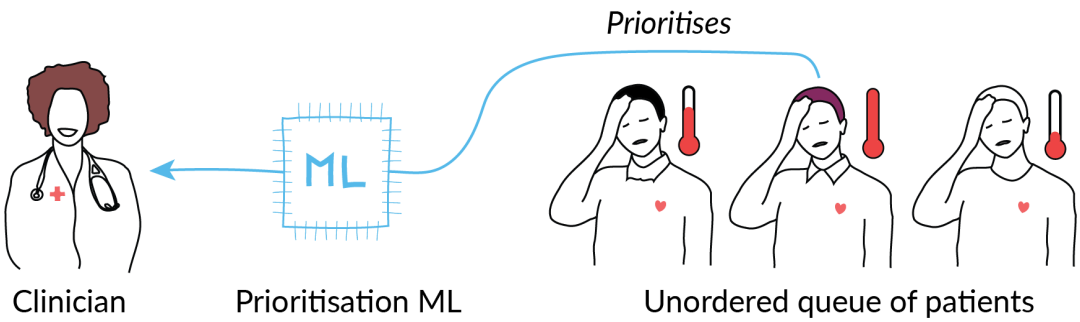


Figure 10: A clinician-facing ML-based system conducts prioritisation of patients requiring attention.

Table 10: Descriptions and examples of the intended use: prioritisation.

PRIORITISATION
Only prioritisation Baxter et al. [25] evaluated a clinician-facing ML-based system that provided a readmission risk score for all patients. Two of the three reviewed systems that predicted sepsis onset did not provide any additional information except the onset risk [127, 223]. In both cases, the attending staff received a notification and proceeded with standard examinations and care. Benda et al. [28] conducted a study based on a system that solely returned a risk score. However, interviewed physicians were open to considering intervention recommendations.
Prioritisation and explanation Four systems explained their risk scores [22, 171, 221, 290, 301, 302]. Besides risk prediction, healthcare providers could use their expertise to judge the prediction or adjust their actions based on the supplemented explanations of the risk scores. The systems offered, among other things, a list of the most contributing factors, e.g., test results, medications, demographic data, and historical diagnoses.
Prioritisation and recommendation Cho and Jin [80], and Romero-Brufau et al. [285] described two systems that, in addition to risk scores of potential adverse effects, returned a list of recommendations on how to mitigate them.
Prioritisation and explanation and recommendation Only one of the reviewed articles reported on a system that explained its risk prediction and provided recommendations on how to address them [286].

12.3.3.3 *Automation*

When autonomous agency and authority are delegated to ML-based systems, they may substitute instead of support medical providers in clinical tasks of the healthcare delivery process (Figure 11). Among the reviewed articles, only one system was described as being delegated such autonomy. Beede et al. [26] implemented a deep learning system capable of assessing retinal images and issuing binding ophthalmologist referrals. Before the ML introduction, retinal images taken at a local healthcare centre were sent for evaluation by a specialised clinician who decided whether to refer a patient to an ophthalmologist. The clinician-facing ML-based system was able to conduct such an assessment autonomously by detecting and determining the severity of diabetic retinopathy, and issuing referrals. This functionality, aimed at reducing the waiting time for patients, effectively automated the work conducted by the specialised clinician.

12.3.3.4 *Positive side effects of ML integration*

The described *intended uses* did not account for all the effects ML-based systems had on medical professionals and their work. ML had a much more profound influence on work practices. We examined

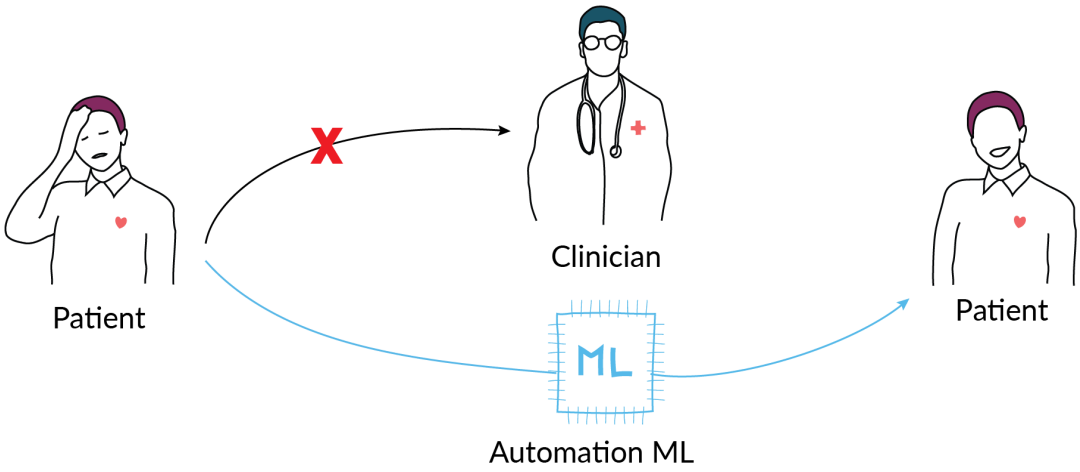


Figure 11: A clinician-facing ML-based system is responsible for a single task of the healthcare delivery process.

three accounts of positive unintended consequences that stemmed from the use of ML.

First, ML served as a discussion catalyst and communication booster [127], which the care team perceived positively. Multiple clinicians emphasised the collaborative nature of their work. Whereas a system may be designed to support a single clinician at the time, in reality, clinicians primarily work as a team [28]. The exact role of the clinician-facing ML-based system within these teams differed depending on the context. However, we discovered a few recurring themes. Clinicians used ML output as a discussion starter about patients highlighted by the system that needed attention [134, 285, 286]. Benda et al. [28] noted that such output could help when discussing requests for additional resources. ML could also support an ongoing discussion. Such output could add gravity to otherwise ignored points raised by staff members with less expertise, e.g., nurses or junior physicians [26, 127, 356] or when trying to convince a patient [26, 221].

Second, the articles reported on a learning opportunity when interacting with ML – an effect that was especially prominent in the decision support *intended use*. Systems supporting case-based reasoning were considered a means to transfer domain knowledge from more-experienced to less-experienced clinicians [134, 172, 334]. Additionally, inexperienced medical professionals used ML's prioritisation and contributing factors to develop expertise and learn about the focal conditions [22, 80, 221, 302]. Clinicians also hypothesised that exploiting new forms of data visualisation provided by the ML-based system could have research applications [237].

Third, the authors reported increased vigilance of medical professionals. In two of the studies targeting sepsis risk prediction [127, 290], clinicians argued that the presence of an ML-based prioritisation system made them more aware of the risk [301] and increased their monitoring [127].

AUTHOR		GOAL OF THE STUDY	P	N	M	NM	O	IT
Barda et al. [22]	C	Designing AI explanations	12	8				
Baxter et al. [25]	C	Identifying barriers to AI utilisation	7	3	3	2		
Beede et al. [26]	I	System deployment and observation		18		1		
Benda et al. [28]	C	Identifying facilitators, challenges, and recommendations		15		25		7
Brennan et al. [57]	C	Prospective evaluation of usability and accuracy	20					
Cai et al. [68]	C	Gathering pre-deployment information needs	12					
Cai et al. [67]	C	Designing interactions	21					
Cho et al. [80]	C	Prospective qualitative evaluation		18				
Gastounioti et al. [122]	C	Design, development, deployment	5					
Ginestra et al. [127]	C	Post deployment evaluation	44	43				
Gu et al. [134]	C	Development	12					
Hollander et al. [155]	C	Pre-post deployment evaluation	27					
Jauk et al. [171]	I	Development and deployment	5	5				5
Jin et al. [172]	C	Design and development	11				1	
Matthiesen et al. [221]	C	Near-live feasibility and qualitative evaluation	8					
McCoy et al. [223]	I	Development and deployment	x			x		
Morrison et al. [237]	C	Visualising AI output	5					
Petitgand et al. [261]	C	Identifying post-deployment adoption barriers	7	3		5		5
Romero-Brufau et al. [286]	C	Pre-post deployment evaluation	x	x	x			
Romero-Brufau et al. [285]	I	Design, development, and deployment	7	21	3		6	
Sandhu et al. [290]	C	Investigating factors influencing integration	7	8				
Sendak et al. [301]	I	Design, development, and deployment	x	x		x	x	x
Sendak et al. [302]	I	Design, development, and deployment	x	x		x	x	x
Wang et al. [334]	C	Post deployment evaluation	22					
Yang et al. [356]	C	Design	23	x				

Figure 12: Overview of study goals (divided between Confined (C) and Innovation (I) types) and involved stakeholders.
P - Physicians, N - Nurses, M - Other Medical Professionals, NM - Non-medical professionals, O - Other, IT - IT specialists. Underlined numbers represent the end users of a given system. X denotes an unspecified number of involved stakeholders. The darker the colour of the background the more people were involved in relation to other studies.

12.3.4 *Stakeholders involved in medical ML innovation*

The reviewed studies presented a range of different goals. To meet them, diverse groups of stakeholders and clinical end-users were involved in various ways. To support researchers and practitioners from the HCI, ML, and Health domains in navigating medical ML innovation processes, we explored the situated approaches characterised by the heterogeneity of involved stakeholders and employed methods. We compiled an overview (Table 12) of the studies' reported goals and stakeholders (divided into six professional groups). We specified evaluations conducted in the wild as a distinct study goal in order to highlight post-deployment evaluation as opposed to evaluations completed during design, development, and deployment.

Among the reviewed studies, we distinguished two major types. First, 19 confined studies explored a particular aspect of medical ML

innovation, e.g., identifying general facilitators and challenges for a particular vision of a system [28] or gathering pre-deployment information needs [68]. Second, six studies described an innovation process, i.e. covering some steps of design, development, and deployment and bringing their system into the wild [26, 171, 223, 286, 301, 302].

Innovation studies involved more diverse stakeholders compared to confined ones. To gain a better overview of the types of stakeholders that were involved, we divided them into six groups based on their position and background, reported their numbers, and highlighted which groups were the intended end users of the systems. While all of the studies involved relevant end-users and stakeholders, they differed in the diversity of their participants. The confined studies mostly involved one to two professions and focused primarily on end-users. By comparison, the innovation studies involved a more diverse group of stakeholders. On average, representatives of 3.5 groups were involved in these studies. It suggests that while it is possible to study certain aspects of medical ML innovation, realising ML in a clinical context is a joint effort of representatives from many different domains.

Physicians and nurses were the most involved groups, with physicians involved in all but three studies. Other medical professionals included, e.g., radiographers or technicians, who were a part of only three studies[26, 285, 286]. Non-medical professionals served as administrative workers - secretaries, clerks, administration [223], managers and board members [28]. Other stakeholders were people who did not fit into any other category, e.g., statisticians [302], designers [172, 261, 301, 302], or self-reported as "other position" [285]. Lastly, some of the studies involved IT specialists - hospital's information officers [28], ML engineers [171], other professionals [171, 301], developers [261, 301], data scientists, and engineers [302]. The diversity of involved stakeholders spotlights the breadth of medical ML innovation projects and foreshadows potential challenges that may arise from it.

12.3.5 *Activities and techniques of medical ML innovation*

To guide medical ML innovation processes and learn from existing studies, we analysed the modes in which stakeholders, including clinical end-users, participated. Overall, we found that stakeholder participation differed across activities and input from stakeholders was rather diverse. We conceptualised 10 unique activities that correspond to 10 distinct goals described in the articles. We did this to accommodate for the high degree of situated differences, the lack of clarity of some of the reports, the interrelatedness of techniques and goals, and the varying order of activities in the medical ML innovation process. Despite the non-standardised descriptions of the reported activities and their varying order of application, the studies provided attainable and well-defined goals. Focusing on the goals

allowed us to progress from fixed stages to, often parallel, activities. Although the order in which we describe the activities may suggest a sequential process, the activities were often overlapping or applied across phases of design, development, and deployment. Progressing from phases to activities enabled us to recognise the complexity and interrelatedness of activities in the medical ML innovation processes. The concrete deconstruction of the activities, employed techniques, and involved stakeholders can be seen in Table 11.

Table 11: 10 synthesised activities of medical ML innovation processes derived from both innovation and confined studies. Each activity includes its goal, employed techniques, and involved stakeholders.

PROBLEM FRAMING	
Goal	Framing the problem space based on a situated and interdisciplinary perspective. Expanding understanding of the focal problem beyond the superficial causes.
Techniques	*Interview:* "informing the implementation of a predictive algorithm" [28] "discussed challenges ... when making diagnosis" [172] "understand complexities in addressing the problem" [302] *In-situ observation:* "Observe frontline staff in clinical settings where the problem occurs" [301] *Undefined collaboration:* "assembling [a] team around the problem," "front-line physicians [...] work with a local innovation team to improve detection and treatment of sepsis" [301] *Data analysis:* "curating the data to better characterize the problem" [301, 302]
Involvement	*End-users:* Physicians [28, 172, 301, 302], Nurses [301] *Other stakeholders:* IT specialists [28, 301, 302], Non-medical Professionals [28]
UNDERSTANDING CURRENT PRACTICES	
Goal	Understanding current work practices and mapping used technologies. It involved learning about future end-users' goals, struggles, and motivations, as well as locating areas that could benefit from ML support.
Techniques	*Interview:* "questions pertaining to current readmission workflows" [25], "15 hours of interviews" [26], "understand how the decision making process ... unfolds" [356] (in [357]) *In-situ observation:* " to ... understand the eye screening process" [26], "to understand the clinical workload" [221], "user-centered observations, workflow-mapping ... were conducted to determine how best to incorporate the tool into existing workflows" [285], "how an implant decision is reached across many clinician roles and contexts" [356] (described in [357]) *Undefined collaboration:* "collaborated ... to map local workflows" [286], "working with front-line clinicians to understand ... the care delivery process" [302] *Undefined work:* "assessments ... to determine how best to incorporate the tool into existing workflows" [285]
Involvement	*End-users:* Physicians [25, 221, 285, 302, 356], Nurses [25, 26, 285, 302], Other Medical Professionals [25, 285], Non-medical Professionals [25] *Other stakeholders:* Undefined [286]
DEFINING NEEDS AND REQUIREMENTS	

Table 11

Goal	Devising high-level requirements for the new ML system. Clarifying assumptions and shaping the future work direction.
Techniques	*Interview:* "to understand pathologists' needs" [67] "they expressed a desire for ... they wished the tool to..." [172] *Mock-up experimenting:* "paper prototypes ... to understand pathologists' needs" [67] *Workshop:* "to understand the information ... they ... use in their decision-making.", "[rapidly sketch their ideas] ... to discuss and clarify requirements." [237] *Undefined collaboration:* "requirements ... specified in collaboration" [122] "invested ... effort into gathering requirements" [301]
Involvement	*End-users:* Physicians [67, 122, 172, 237], Nurses [301] *Other stakeholders:* Physicians, Other Medical Professionals, Other, IT specialists [301]
SYSTEM DESIGN	
Goal	Refining, clarifying requirements, materialising future solutions, and developing a working vision. A platform for discussion and early testing of the future system.
Techniques	*Workshop:* "a guided group review and critique of ... prototypes" [22], "repeatedly met with ... nurses to iterate on functions, information, control, and visual components of the design" [301], "co-design ... focusing on ... sketching the user interface" [221] *Prototyping:* "created ... functional prototypes and iterated on [them] with further feedback" [67], "an interactive ... prototype ... was demonstrated to ... refine the initial requirements" [172], "presenting the design ... opportunity to try the app" [237] *Design feedback:* "present a set of visualisations" [237], "iterated on the design based on feedback" [356] *Interview:* "we informally interviewed nine neurologists and asked them to discuss three potential visualisations" [237] *Survey:* "a questionnaire to indicate preferred design options" [22]
Involvement	*End-users:* Physicians [22, 67, 172, 221, 237, 356], Nurses [22, 301]
NEW WORKFLOW DESIGN	
Goal	Conceptualising the future work practice with the new system as an integral part of the new workflow.
Techniques	*Undefined collaboration:* "interdisciplinary team ... designed a workflow" [290], "collaboration ... to finalize workflow decisions" [301], "collaborated ... to identify how to ... integrate the tool" [286], "a transdisciplinary team ... designed the ... workflow." [302] *Team decision:* "To situate a DST into the current VAD decision-making routine ... we chose the ... meetings" [356]
Involvement	*End-users:* Physicians [286, 356], Nurses [290, 301, 302] *Other stakeholders:* Physicians, Non-medical professionals, IT specialists [290, 302]
ML MODEL DEVELOPMENT	
Goal	Sensitising development team to domain knowledge and informing the ML model development.

Table 11

Techniques	*Interview:* "probes regarding ... the model including external rules and regulations, internal organization, clinical content" [28], "to enumerate diagnostically important concepts" [67], *Workshop:* "Feature engineering was carried out ... during five co-design workshops." [221] *Focus groups:* "feedback ... reported to the ... developers during focus groups" [261] *Previous implementations:* "sensitivity ... was determined a priori and informed by our experience with EWS 1.0" [127] *Delphi method:* "attribute weights elicited from experts using a Delphi method" [134] *Data-focused collaboration:* "[data-generating examinations] were chosen by our MS expert clinicians" [237] (in [192]), "Clinical experts specified and reviewed .. all [the used] data" [302], "the department of nursing [requested removing a data point]" [80]
Involvement	*End-users:* Physicians [28, 67, 134, 221, 237, 261, 302], Nurses [80, 261, 302] *Other stakeholders:* Non-medical professionals, IT specialists [28]
SYSTEM DEVELOPMENT	
Goal	Realising the complete system through progressing from concepts and ideas to an operational system. Includes parts of solution conceptualisation, new workflow design, and ML algorithm development.
Techniques	*Pilot study:* "to assess the usability and accuracy ... using a simulated workflow" [57], "comments ... during ... expert group meetings before and during the pilot", "the expert group suggested improvements ... and new functionalities ..." [171] *Iterative development:* "cycles to evaluate processes and incorporate clinical feedback" [223], "iterations that explore the best way to communicate visually sensed data" [237], "feedback loops were crucial to improving Sepsis Watch" [302], "reviewed multiple versions of the user interface", "clinicians specified the most relevant information to accompany the risk level" [301] *Case study:* "to validate ... the refinement mechanisms", "how SMILY ... affects user experience and search practices during a medical task." [68], "provide evidence [of] the usefulness of the system." [172] *Interview:* "usefulness, ease of use, general pros and cons of the prototype system, visualization designs, and insights" [172] *Survey:* "regarding the usability of the algorithm and web interface" [57]
Involvement	*End-users:* Physicians [57, 68, 171, 172, 223, 237], Nurses [171, 302] *Other stakeholders:* Non-medical professionals [223]
DEPLOYMENT PREPARATION	
Goal	Anchoring a vision of the new system within the organisation, informing end-users about the upcoming changes, training affected staff in the use of the new system and preparing the technical infrastructure for the eventual deployment.

Table 11

Techniques	*Training session:* "training necessary for effective human-AI collaboration" [68], "guidance on how to use [the system]" [80], "various training sessions" [171], "education sessions to train ... on the proper uses" [223], "in the program workflow and application" [290], "one-on-one training sessions" [301] *Training materials:* "preparing training materials" [286], "collaboration ... to develop training material", "A "Model Facts" sheet" [301] *Promotion:* "via emails" [127], "promoted the application throughout all participating departments" [171], "in faculty meetings and via email" [290, 301] *Undisclosed training:* "training staff" [286], "instructed to consider the list of recommendations" [285], "educated on the model's aggregate performance measures" [290]
Involvement	*End-users:* Nurses [68, 80, 127, 171, 285, 290, 301], Physicians [127, 171, 223, 285, 290]
DEPLOYMENT	
Goal	Physical installation of the system, transitioning to the new workflow. Supporting affected staff members, solving problems encountered after increased real-world usage, and assessing the new system's effects.
Techniques	*Continuous feedback:* "weekly feedback phone calls" [26], "quality improvement team met regularly" [223], "formal and informal lines of communication" [301] *Gradual deployment:* "some oncologists have begun to use the system on trial" [134], "During this cycle ... implemented into ... ED" [223], "rolled out to a small group of RRT nurses" [301] Parallel workflows: "use of the [ML] scores ... [and] continuing standard procedure" [223], "A 3-month silent period" [302]
Involvement	*End-users:* Physicians [134], Nurses [26, 223, 301] *Other s'takeholders:* IT specialists [301]
EVALUATION	
Goal	Measuring staff's acceptance level, sentiment, and actual use of the new system. Although it is not strictly an improvement initiative, it includes collecting feedback for future developments. Usually the final activity of a development process.
Techniques	*Survey:* "The first survey ... after launching the service", "The second survey ... 9 months later." [80], "user acceptance ... using questionnaires seven months after implementation" [171], "end-user satisfaction ... questionnaire" [122], "web-based questionnaires to assess clinician perceptions" [127], "a survey of the users" [134], "[a survey] after completion of the trial" [155], " user acceptance ... using questionnaires seven months after implementation" [171], "the post intervention survey" [285] *In-situ observation:* "Observations focused on ... the use of systems" [334] *Interview:* "The interviews focused on ... user experience and perceptions of [the ML-based system]" [334]
Involvement	*End-users:* Physicians [122, 127, 134, 155, 171, 285, 334], Nurses [80, 127, 171, 285] *Other stakeholders:* Other [285]

Similar to the conceptualisation of activities, the accounts of techniques varied in naming conventions and description styles. To provide a more general overview, we interpreted the use of tools and

techniques and grouped them into categories. The authors' descriptions differed in their degree of detail. The descriptions that were more abstract were categorised as *undefined collaboration*, e.g., *"collaborated ... to identify how to ... integrate the tool"* [286]. Perhaps due to their goals or varying target audiences, some articles offered in-depth descriptions, while others were limited to a few words about the essence of their technique, e.g. *"[used] paper prototypes ... to understand pathologists' needs"* [67] vs *"requirements ... specified in collaboration"* [122], which were reported in two confined studies, in the *defining needs* and *requirements* activity. Interviews were the most used technique - reported in 7 of the activities. The remaining activities that did not incorporate it were new workflow design, *deployment preparation*, and *deployment*. The second most used technique was workshops - presented throughout three activities: *defining needs and requirements, system design,* and *ML algorithm development*.

While undefined or informal techniques were present throughout most of the studies, they were used to a greater extent in the articles describing innovation studies. This could be linked to the fact that the innovation studies more closely resembled real-world innovation processes, and involved more diverse groups of stakeholders. We observed that undefined collaboration was present in the following activities: *problem framing, understanding current work practices, defining needs and requirements*, and *new workflow design*. Most of these activities involved diverse groups of stakeholders. Moreover, they took place before a coherent vision for the future system was ready. At the time of their execution, there were no complete systems, realised designs, or clinical decisions, which may have imposed more abstract modes of collaboration. This suggests that there may be fewer disseminated methods and techniques for collaboration during medical ML innovation.

12.3.6 *Sociotechnical Challenges*

In this section, we present five sociotechnical challenges that emerged from the studies. As a way to disambiguate the intimate connections between the technical and social, we discuss five interdependencies that emerge as particular challenges for the innovation of clinician-facing ML-based systems. We can analytically distinguish between three technical and three social areas where problems with the ML-based medical systems were rooted (Figure 13). The technical areas were (i) training data & ML model, (ii) system integration & data used, and (iii) user interface. The social areas comprised (iv) users & system use, (v) workflow & organisation, and (vi) healthcare institution & political arenas. While problems may be categorised according to where they are rooted, we found that challenges emerge from being sociotechnically constituted and cannot be analysed detached from their context. These relationships suggest that using measures that tackle only one of the aspects may not be enough and that more comprehensive actions during the ML innovation process may be needed.

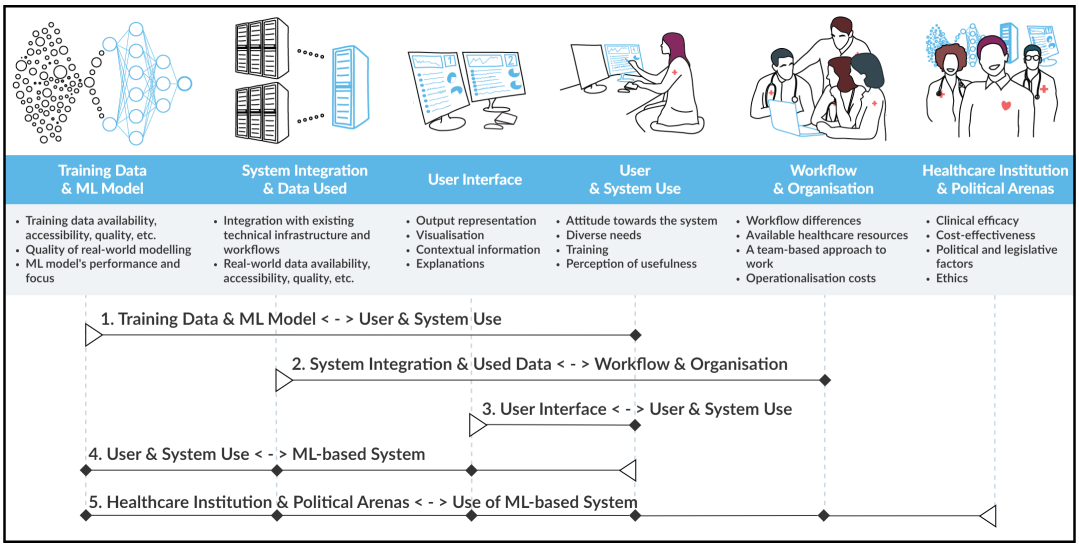


Figure 13: Five sociotechnical interdependencies of clinician-facing ML-based systems' innovation.

12.3.6.1 *Training Data & ML Model <—> User & System Use*

Poor training data quality and inadequate ML models were described as causing challenges during clinical use. Across the studies, we found **three characteristics of poor quality training data: quantity, consistency, and comprehensiveness.** For example, clinicians remarked that *"the quality of the EHRs collected in Chinese hospitals were much worse than those of the MIMIC dataset [an open data set]"* [172]. They explained that the low quality of local training data was due to missing and inconsistent data points. Moreover, because clinics were only starting to use digital EHRs, not all of the data was digitalised, which reduced the possibility of longitudinal analysis. Interestingly, even when the data was recorded, access to codified, quantified, and structured data could be limited, as described by Romero-Brufau et al. [286]. Lastly, **challenges arose when some of the data were not considered during modelling** e.g. patient individual information [334], or social stressors like unemployment or loss of a family member [25, 285].

The above examples resulted in modelling issues, poor precision and alignment with clinical reasoning, which negatively affected how useful clinical end-users perceived the systems. Weakness in modelling the complexities of healthcare work manifested as **unduly general and simplistic recommendations that did not provide new insights**. These complaints were described as, e.g., *"the physicians interviewed considered that the AI-based system was ... very poor at making sense of multicomplaint conditions (pain throughout the body, pain related to severe pre-existing conditions, etc.)."* [261] or returning not useful information like *"a high-risk prediction for a sedated and paralysed patient"* [22]. In such cases, the ML output not only led to no change in clinical action [127] but also connoted that ML does not "understand" the clinician's job [285] and undermined its perception of usefulness [22]. Several studies also described issues with the low accuracy of the

ML models. For example, inaccurate classification of standard-risk patients as high-risk patients led to dissatisfaction among health professionals in primary care [285] and low accuracy in sepsis detection led to significant alarm fatigue [301, 302].

These findings indicate that technical factors like training data and ML models are highly bound to the social or human elements, such as clinicians' experiences. For example, inconsistent training data generated "concerns", and inadequate models led to "confusion". This means that **training data and machine learning models need to be understood as interdependent with clinical end-users and their use of the ML-based system.** In other words, training data and models may have seemed promising in the lab. Yet, when exposed to clinicians in real-world situations or simulated workflows, it became overt whether the quality of the training data was good or bad and whether the ML models were adequate or not.

12.3.6.2 *System Integration & Data Used <—> Workflow & Organisation*

While challenges may be rooted in training data and ML modelling, issues also arose from systems' integration and real-world data available after deployment. Several studies described how these technical challenges negatively affected local clinical workflows. Across the studies, we distinguished four types of system integration issues and real-world system performance that affected the flow of work.

First, issues with **integration of the ML-based system into the existing suite of technological solutions** [80, 122, 134, 261, 285, 334, 356]. Second, **integration issues affected quality of the available data**, e.g., a system investigated by Wang et al. [334] lacked a connection to the local pharmaceutical system, which resulted in the use of outdated prescriptions. Similarly, Matthiesen et al. [221] reported on possible structural data differences on a patient level, i.e., medical devices feeding data to the AI-based system could have been configured to detect and transmit only certain events. Third, **the real-world data affected ML performance**, which caused challenges to the workflows and organisations [22, 26, 285, 301, 302] e.g. algorithm using real-world data significantly increased the number of false positives [301, 302] or real-world images could not be graded in 21% of all cases [26]. Lastly, **untimely delivery of ML output decreased its usefulness**. Several studies described temporal issues afflicting system integration. For example, poor timing was an issue that manifested through notifications arriving "too late" [28, 155], or pertaining to an already diagnosed patient [127], which was perceived as irrelevant "in-the-moment" of the current workflow [67].

The above-mentioned **integration issues had several repercussions on the flow of work at the clinical sites.** In the study by Beede et al. [26], the real-world integration introduced a new image quality check that increased the average time needed to screen one patient. Local contextual issues like poor lighting conditions, which always had been a factor for nurses taking photos, *"but only through using the deep learning system did it present a real problem, leading to ungradable*

images and user frustration". For example, the availability of dedicated rooms, inconsistent screening procedures, and a high number of patients had consequences on how useful the underlying ML model and system were found. Similarly, Sandhu et al.'s [290] account of how physicians in the emergency department had **difficulties with aligning the team-based approach** (physicians, nurses, and residents) with sharing new information coming from the ML output and system: *"For me to have to track them both down to give them that information would be burdensome and that's what would get in the way of flow in the [emergency department]"*.

Moreover, issues of poor performance on real-world data related directly to increased workload and being overburdened. Medical professionals were often concerned about **the additional time that was necessary to operationalise the ML-based systems.** In the world of understaffed and underfunded clinics, spending additional time on data entry [285, 334], overthinking [67] or outright pursuing wrong diagnoses suggested by the system [261], waiting for ML output [122], or filtering through a sea of false-positive alerts [301, 302] completely disrupted so-far functioning work practices, and shifted clinician's attention from patients to the systems. Clinicians from two studies warned against **an over-utilisation of healthcare resources** [28] and painted a picture where not all of the required bedside assessments would be possible [223]. The additional costs of complying with recommended clinical actions were of concern to nurses in the study by Cho et al. [80], who observed that some of the recommendations by the ML-based systems were deliberately ignored.

These findings suggest that while the underlying ML model and system may prove to have promising performance in laboratory settings, success in clinical practice hinges on the quality of integration, interoperability with other IT systems, and access to good-quality data when the ML-based systems are deployed. This means that the technical aspects of the ML-based system needs to be considered as interdependent with workflows and real-world contextual factors like clinical staff, build environments, and monetary resources.

12.3.6.3 *User Interface <—> User & System Use*

The third technical area from where problems emerged was the user interface of an ML-based system. More than two-thirds of the studies described critical issues with the user interfaces and the interpretability of the ML output. Among them, we identified four main types of problems. First, **missing or poor explanations** and failing to present reasons for the ML-based output negatively affected the use of the systems. For example, exposing intricacies of algorithmic decision-making that were incomprehensible to clinicians [237], triggering ML-based alerts of severe sepsis and not providing reasons for the result [127], or presenting predictions of diagnosis without revealing the impact of historic data [172] created issues with the utilisation of the output. Second, we observed that **interactive ML models could be too captivating**. For example, some clinicians interacted with the ML-

based systems by choosing relevant features and receiving updated predictions to test their hypotheses [22, 67, 127, 172], however, some clinicians cautioned against *"going too deep down a tangential rabbit hole"* [67]. Third, **missing contextual patient information alongside the ML output** surfaced as a critical issue in several cases e.g. when minimal patient information was missing for assessing the clinical relevance of a prediction [22] or when the ML-output alone was not enough to support clinical action [290]. Fourth, **poor presentation of the ML output in the user interface** created interpretation issues. The issues occurred when presenting: multiple risk scores at the same time [25]; risk predictions in exact percentages instead of broader risk categories [221]; or too many confounding features [67]. Providing clear, understandable, clinically relevant, and actionable output was regarded as essential for utilising the system's output [28, 67, 171, 221, 237].

Such issues with interpretability, interactivity, and presentation affected clinicians and the ways they used the systems. **Physicians' trust in the system deteriorated due to poor explainability and black-box issues**, e.g., *"[w]e sometimes hesitate to trust the machine learning models because they usually fail in providing reasons"* [172]. The distrust was deepened by not displaying contextual patient information. Clinicians reported the need for more individual and cohort information to assess the relevance and to trust the predictions: *"Historical events can provide evidence for us to determine whether the prediction is trustworthy"* [172]. Similarly, nurses that received a sepsis alert were uncomfortable making decisions based on the minimal patient information available [290], and electrophysiologists' interpretation of risk predictions were, in some cases, dependent on contacting the patient for more information [221]. In paediatrics, clinicians lacked patient information like current disease state or baseline risk to assess the clinical relevance of the predictions [22]. In this way, the interpretability of the algorithmic ML output was, for many clinicians, dependent on the contextualisation of the model's classification next to relevant patient information. Moreover, clinicians' intuition was hampered, as described in one case where the display of important variables, not previously associated with the outcomes, led to poor interpretability [127]. **Clinical accountability was also found to be affected by insufficient explainability**, e.g., Sandhu et al. [290] reported how nurses took on a responsibility to explain the ML risk score to physicians, which created a mismatch in understanding and disturbed the nurse-physician relationships. Similarly, Cho et al. [80] described how nurses' accountability was affected by interaction with the ML model through data entry - *"nurses said that they became very careful about documentation due to the thoughts that the data they entered will be used to infer risks of falling"*. Providing ways for interacting with the underlying ML model was sought-after [67, 172] but also led to possible confirmation bias as described in one case: *"If I'm adjusting that bar; [...] I'm injecting too much of my interpretation into it, how much of this is me putting in my subjective interpretation hoping to get that response back?"*. Finally, when the visual characteristics of the ML-based user

interface were unclear, clinicians found it to generate "confusion" [25], muddy the results [67], or make it "hard to translate" into clinically relevance [221].

The interdependencies between the context of use and ML output visualisation, contextualisation, and interactivity, suggest that sociotechnical challenges emerge from the interplay between the ML-infused user interface and the clinical end user. This means that the visualisations of ML output and the user interface need to be conceptualised, designed, and developed in connection to each other.

12.3.6.4 *User & System Use <—> ML-based System*

Issues emerging from the misalignment between clinical end-users and their use of ML-based systems were described in more than half of the publications. We found three human and user-related factors that affected the experience of the capabilities and limitations of the clinician-facing ML-based systems.

First, **clinicians general attitudes and feelings about machine learning** influenced the experience of the performance and usefulness of the overall ML-based system: “Fear of overstepping”, “feeling uncomfortable”, "resistance to change"[290], and "sceptical attitudes"[261]. In some cases, physicians felt their knowledge of patients’ histories and circumstances was more accurate than risk score estimations generated by the system [25, 127]. Some clinicians reported that peer-reviewed publications on an ML-based system’s performance were the preferred form of reporting and that such publications would increase their trust in the system [221, 356]. However, when physicians were too optimistic about what ML can do, and when they had too high expectations about the capabilities of ML-based predictions, the perceived usefulness of the system was affected negatively [221].

Second, **diverse clinical roles and diverse needs** affected the perceived usefulness of the overall ML-based systems. Several studies highlighted that systems which engaged multiple clinical specialities had to conform to mixed preferences. For example, using overall mortality risk predictions [22] in paediatrics, nurses wanted minimal, actionable information and generally preferred simple and static explanations. On the other hand, physicians preferred more dynamic output that they could engage. Similarly, nurses and providers had different needs and perceptions of an ML-based warning system for sepsis prediction: almost half of the nurses found the overall system helpful as opposed to less than a fifth of "providers" [127]. Different needs also arose from differences in the level of expertise, e.g., junior and less experienced nurses were more open to the ML predictions and recommendations than senior and experienced nurses [80].

Third, **lack of end-user training and promotion, unfamiliarity with the ML-based systems, and insufficient computer literacy among clinical end-users** were described as barriers to successful clinical implementation of the ML-based systems. Despite positive feedback about the usefulness of the ML-based system, *”the actual*

system use was low" due to modest outreach and promotion among medical professionals in the department [171]. "Misperception" [127], "misunderstanding" [290] and "confusion" [172] were attributed to the unfamiliarity with machine learning and the lack of training: *"We (doctors) spend years in school to learn how to make [a] diagnosis based on those [traditional] statistical tools and diagrams... your tool is obviously more informative but we just need more time to get familiar with it."* [172]. Lack of knowledge about the capabilities and limitations of the ML tool could also lead to degradation of trust among clinical end-users [68, 285]. When providers had little knowledge about ML and predictive modelling, they found themselves unable to assess and verify the ML model information and credibility [22]. While the absence of training was reported as problematic, some physicians considered trust in the ML-based system as something that emerge over time and from multiple engagements and experiences with actually using the new technology: *"Just like with all other new technology based on machine learning: the first 2 months I sit and read through to see what I have, but in month 3, I will look at the [ML] output alone. Because then I trust that it has pulled out what is appropriate [. . .]".* [221]

These findings suggest that ineffective adoption of ML-based medical systems may be rooted in human and social dimensions related to the clinical end-users rather than in the technical aspects of the ML-based systems. This means that professionally diverse end-users, their attitudes, perspectives, expectations, and training need to be considered as bound to the complete ML-based system. This suggests the existence of important interdependencies between clinical end-users and the system's capabilities and limitations. This has implications for HCI by imposing a strong need for simultaneous configuration of the technical and social areas of clinician-facing ML-based systems and recognising that successful innovation requires iterations or, at best, convergence between phases of design, development, and deployment.

12.3.6.5 *Healthcare Institution & Political Arenas <—> Use of ML-based System*

Issues that were rooted in the context of healthcare institution and political arenas were reported in half of the publications. Four factors stood out as being broader in scope and addressing the wider institutional arenas that included: the political, economic, ethical, and medical professional arenas. These higher-order institutional factors affected the use and perception of ML-based systems in various critical ways.

First, factors related to the **wider medical professional and academic arenas** affected the perceived usefulness of the complete ML-based system. For example, ML provided a novel way of clinically looking at multiple sclerosis. However, there was a mutual agreement among clinicians that it was hard to imagine how it would be of benefit as part of everyday clinical practice [237]. Similarly, issues emerged when, e.g., there was weak consensus on definitions of clinical diag-

nosis [122, 223]; ways of applying clinical guidelines differed across clinical sites [67], or when the unpredictability of the disease in current medical practice was high [237]. These issues suggest that poor clinical utility can arise from the disconnect between existing medical circumstances and the potential and actual use of the ML-based system.

Second, **clinical efficacy and cost-effectiveness** of ML-based interventions were critical for acceptance and adoption. This includes the perceived clinical utility and the clinical outcomes, i.e., the measured effects on patients as a result of an intervention, which were critical factors for acceptance and adoption. Poor clinical outcomes afforded by the ML-based system were problematic and led to non-use in clinical practice. For example, Yang et al. [356] reported that due to low clinical outcomes, merely 7% of the physicians used the neural network score in clinical decision-making. Hollander et al. [155] found that despite success in the lab, the ML-based clinical decision support tool failed to induce a significant improvement in healthcare outcomes. The poor clinical utility hampered the use of the ML-based system. For example, the ML-based clinical decision support tool made no difference in the clinic's effectiveness in reducing complications among diabetic patients [285]. Clinicians considered the ML-based prediction tool as "nice to have" rather than a "need to have" [221, 334], or only half of the physicians deemed the new technology helpful [57]. Moreover, if the use of the ML-based system required actions but insufficient resources were available for carrying out the intervention it led to frustration and growing distrust [28].

Third, **political and legislative factors** were decisive for the success of the complete ML-based system. Benda et al. [28] described a reimbursement system as a core part of the existing clinical reality. As a result, ensuring adherence to external regulation emerged as an issue during deployment in clinical environments.

Fourth, **ethics considerations were also found to affect the overall perceived usefulness** of the ML-based medical systems. Although surprisingly few papers explicitly addressed ethical concerns (8 out of 25 articles), some papers did describe important findings on this matter. In the study by Beede et al. [26], nurses were burdened with weighing the trade-offs and deciding whether or not to enrol patients to be assessed by the evaluated system: *"some nurses felt the need to 'warn' patients that they would need to travel should a referral be given. Given the far distance and inconvenience of getting to Pathum Thani Hospital, 50% of patients at clinic 4 opted out of participating in the study"*.

These four issues demonstrate that problems with realising ML-based systems in clinical settings are bound to broader institutional arenas. This means that the race for getting the technology right in laboratory settings is not enough. There are political, economic, ethical, professional, and academic arenas that are interdependent with the organisational implementation and use of ML-based systems. In order to achieve successful adoption and acceptance with the innovation of clinician-facing ML-based systems, there is an inherent need to attend to wider institutional factors, which are deeply embedded

in the clinical realities and which often challenge the design, development, and deployment of ML-based systems in healthcare.

12.4 DISCUSSION

Design, development, and successful deployment of clinician-facing ML-based systems is a uniquely complex endeavour. This systematic review shows that going beyond confined studies and undertaking a successful medical ML innovation process pose particular methodological challenges for HCI and the interdisciplinary collaboration with ML and Health researchers and practitioners, as well as professionals from wider political and economic arenas. In the following, we first outline the conceptual contribution of this paper. This is followed by a discussion of key findings in relation to existing literature and their implications for HCI. We conclude each section by describing opportunities for HCI and collaboration with Health and ML partners.

12.4.1 *Towards a Conceptualisation of Medical ML Innovation*

Researchers and practitioners from HCI, ML, and Health can use the results of this systematic review as a conceptual framework to base their collaboration on and find a common understanding. We conceptualised four areas that can be relevant in the context of medical ML innovation.

First, we analysed the technical aspects of the systems described in the reviewed articles (see section 12.3.2). Designers' lack of understanding of the technical aspects of ML is known to be a key issue in HCI when designing human-AI interaction [354, 355]. We investigated the influence of technical aspects on the innovation process. The discussed topics included, among others, reliance on domain experts' knowledge, data needs, and explainability potential. The in-the-wild consequences of these technical aspects may help researchers and practitioners from HCI and Health to engage on a more equal footing with their ML partners.

Second, we used the concept of *intended use* to tease out the different purposes and goals of the clinician-facing ML-based systems (see section 12.3.3). While a clear definition of the *intended use* is required to obtain regulatory approval [238], research points out that conceptualising functionalities and future use of ML-based systems is a nontrivial task [104, 355]. Nonetheless, it is imperative for clinical adoption that the ML model is designed and developed with its real-world deployment in mind, which can be achieved by ensuring a robust link between ML and meaningful clinical and operational capabilities [215]. We propose to use the conceptualisation of the *intended uses* to guide the collaborative effort of working and re-working a shared vision continuously and throughout the collaborative activities of design, development, and deployment of the ML-based system in clinical settings.

Third, we conceptualised ten activities and related techniques that were employed in confined studies and during medical ML innovation processes (see section 12.3.5). While the activities do not cover all domain-specific tasks, e.g., data acquisition or clinical trials, the collection of 10 activities may serve as a basis for successfully undertaking the design, development, and deployment of clinician-facing ML-based systems.

Fourth, we characterised five sociotechnical challenges that are particular to medical ML innovation processes, which arise from the interdependencies between the social and technical areas of clinician-facing AI-based systems and their use (see section 12.3.6). Our analysis of the challenges can direct research and development teams towards the non-trivial interrelations that constitute these types of systems and that require special attention during the innovation process.

12.4.2 *Bridging Disciplinary Differences between HCI, ML, and Health*

Research described that achieving collaboration across the disciplines of HCI, ML, and Health is difficult due to epistemological and methodological differences. Blandford et al. [43] contrasted the disciplinary variances between Health and HCI and emphasise the lack of mutual understanding as a grand challenge for research and development of interactive digital health interventions: *"until there is much greater mutual understanding and mutual valuing of the complementary research traditions than exists at present, people risk disappointment and rejection in trying to bridge the divide"*.

The disciplinary differences between HCI and ML are also found to be challenging in ML-based projects. Grudin [132] argued that the HCI and ML paradigms differ so much that they are, historically, contradicting [132]. Similar perspectives are raised in the respective communities [128, 181, 353] and it is recognised that success with ML-based systems requires an extra effort of the ongoing collaboration between Health, ML, and HCI team members [1, 7, 77, 354].

Health research and evidence-based medicine are committed to sequential development processes and randomised control trials [43]. In Health research and development, ML-based medical systems fall under digital health interventions. Historically, based on drug development processes, Medical Research Council in the UK guided the development and evaluation of complex interventions [70]. Their approach can be characterised by its sequential nature, starting with a hypothesis and finishing with a Randomised Control Trial (RCT) that measures the effectiveness of an intervention or a drug. The process is systemic, rigorous, and shielded from external factors [43]. Although new, more flexible approaches have been proposed [83] and new guidelines that suggest iterative development are in place [86], the sequential nature is deeply rooted in healthcare development processes [43].

ML research and development processes are characterised by data work, the mutability of capabilities, and late realisation. ML (as part

of the AI community) and HCI have been portrayed as *"having opposing views of how humans and computer[s] should interact"* [343]. Winograd, similar to Grudin, recognises the historical differences between these two communities. They accentuate the differences in how software should be created, and how interactions should be accounted for. To some extent, the ML community represents the "rationalistic" approach that assumes human actions and thought processes can be *"captured in a formal symbolic representation"*. Such capture is achieved through data, which is one of the main focuses of ML engineers. In a nine-stage iterative ML development process proposed by Amershi et al. [6], three initial stages target data collection, cleaning, and labelling. These stages are collectively called "data wrangling" and account for up to 80% of all the resources spent on data science projects [138, 177, 271]. It comes as no surprise that given such disparity, data and data-centric approaches are at the core of ML work. After all, model evaluation can happen only after extensive data work [6]. The previous steps are characterised by the constant mutability of ML capabilities through retraining, parameter changes, and more data work. Due to the nature of that process, it is impossible to conceptualise all the aspects of ML-based medical systems before their use. Such late realisation means that final capabilities take shape only after the ML-based system's deployment [128, 354].

HCI is committed to understanding users and the context of use. HCI researchers and practitioners have developed principles, overall methods, and techniques that focus on mutual learning and are necessary to foster a collaborative development process and meaningful involvement of stakeholders. There has been a longstanding interest of the HCI community in engaging with software engineering and practice [133]. Several methodologies aiming to incorporate HCI perspectives into and support the development process have been proposed throughout the years [297]. HCI also offers a closer look into collaboration during such processes. Piorkowski et al. [265] highlight three communication challenges in interdisciplinary environments. The major themes centre on "knowledge gaps across roles", "establishing trust", and "setting expectations". Drawing from the HCI theory, researchers and practitioners are in a position to foster collaborative interdisciplinary development processes.

12.4.2.1 *Success with medical ML-based innovation hinges on interdisciplinary approaches and extended stakeholder involvement*

In this systematic review, we found that machine learning and the related technical choices affect the requirements for multi-disciplinary expertise and collaboration between HCI, ML, and Health researchers and practitioners. Moreover, innovation studies increasingly engage a diversity of stakeholders and apply informal or unspecified techniques.

ML affected collaboration among end-users and other stakeholders. The choice of the ML algorithm determined the dependency on domain input (see section 12.3.2). Knowledge-driven algorithms re-

quired in-depth collaboration on feature engineering [221]. By contrast, data-driven algorithms that derived features from annotated data [203] required less collaboration and afforded fewer opportunities for mutual learning and the *ML development approach* shaped the collaboration space between the partners from HCI, ML, and Health, e.g., Beede et al. [26] used open data sets to *discretely* create their ML model, and the collaboration with domain experts was limited to providing annotations [137]. Moreover, ML-based medical systems delivered by third-party providers limited the opportunities for mutual learning, especially during the initial activities, e.g., [223, 285].

Innovation studies applied additional informal techniques and engaged a diverse set of stakeholders. Innovation studies applied well-cited HCI techniques like user-centred observations [285] and interviews [302], however, the same studies also applied unspecified, informal techniques using formulations like *"working with frontline clinicians"* [302], *"assessments [...] to determine"* [285], and *"cycles to evaluate processes and incorporate clinical feedback"* [223]. On the contrary, confined studies tended to apply only well-established techniques like interviews [25, 28, 172] and prototyping [67, 172, 237]. These findings suggest that the process of successfully transitioning clinician-facing ML-based systems into medical settings, to some extent, will have to escape traditional domain-specific techniques. Innovation studies also involved a more diverse set of stakeholders. While most studies in the review included clinical end-users, innovation studies stood out by engaging other medical professionals, administrative personnel (e.g. secretaries, clerks, administration), managers and board members, and IT specialists (e.g. hospital's information officers, software developers). This means that while it is beneficial to engage clinical end-users, the lab-to-clinic transition requires extensive collaboration between a broader range of professional expertise.

12.4.2.2 *Implication for HCI: Need for interdisciplinary collaboration and striving for mutual learning*

Fostering meaningful collaboration and aligning stakeholders from various domains with different traditions and values is historically a part of HCI's agenda. Researchers from the Computer Supported Cooperative Work (CSCW) and Participatory Design (PD) domains developed methods for close collaboration and balanced software development [185, 249]. However, as presented in this review, the technical possibilities are realised late in the ML innovation projects, and best practices are often developed within separate domains. Moreover, ML poses additional challenges to the innovation processes, like shaping the interaction space or requiring new forms of interdisciplinary engagement. With these challenges at play, it is pivotal for the collaboration to create shared understandings and to foster mutual learning between the stakeholders, researchers, and practitioners. Despite these challenges, the HCI community is particularly well-suited to support such interdisciplinary and uncertain collaboration. In particular, PD offers principles, tools, and techniques to shift power to end-users,

while considering organisational goals, work practices, and changes that need to follow [187, 306].

Opportunity #1: Focus on the joint demystification of ML when long-term engagements are impossible. Innovation of clinician-facing ML-based systems oftentimes involves stakeholders whose participation in the project is only a fraction of their daily responsibilities [25, 28, 57, 221, 261]. In such cases, costly and time-consuming engagements proposed by PD, while fruitful in the long run, may not always be feasible. Instead, researchers have explored new methods of collaborative demystification of ML. Yu et al. [358] concentrated on increasing designers' and end-users understanding of the trade-offs between design objectives and helping them navigate the model selection. The developed method employed a visualisation tool that translated the trade-offs into the end-user's practice. Another group has focused on alleviating the challenge of using AI as a design material [104, 181]. Subramonyam et al. [316] propose using end-user data as a data probe to facilitate *"divergent thinking, material testing, and design validation"*. The pairwise collaboration of User Experience (UX) designers and AI engineers informed future designs, as well as the design of "AI architecture". Further, methods for improving user experiences and expectations have been centred on the problem of collaboration between HCI designers, ML engineers, and end-users (see e.g. [352, 354]). The uncertainty of possibilities and constraints of ML is one of the core design challenges that should be tackled by intentional interdisciplinary collaboration.

Opportunity #2: HCI researchers and practitioners should step out of the comfort of established and contained techniques. Mitigating the interdependent challenges of medical ML innovation, as well as supporting profound interdisciplinary collaboration require intentional effort. As pointed out by Bødker et al. [49], collaboration and mutual learning during software development are too rich and nuanced to be captured by a single technique. Instead, there is space for building relationships and mutual understanding during collaboration through "participatory infrastructuring" before, in between, and after the execution of conventional techniques. Similar flexibility and openness in creating design spaces were advocated by Bjørn et al. [40], who underscore the agency of HCI and CSCW researchers and practitioners in the design of collaborative space-time, in contrast to serving solely as providers of appropriate techniques. Such spaces could serve a purpose of a "third space" [241], which is a space not "owned" by HCI collaborators or used solely to elicit knowledge from participants. Rather, it should be used to negotiate design, exchange perspectives, vocabularies, traditions, and goals, and become a mutual learning opportunity [50]. With all collaborators on a level playing field, such spaces would foster mutual learning and understanding, needed for the successful innovation of clinician-facing ML-based systems. Moreover, HCI researchers' and practitioners' analytical sensibility and deeply rooted interdisciplinarity can yield a better understanding among other parties who often lack collaborative expertise.

12.4.3 *Mitigating Sociotechnical Interdependencies in Medical ML Innovation*

We identified five challenges of clinician-facing ML-based systems, which are characterised by how they emerge as social and technical interdependencies. The conceptualisation of challenges as sociotechnical showed that problems were neither purely technical nor purely social, but an effect of their interaction. Several scholars and more recent studies applied a similar lens and call for sociotechnical approaches to the design, development, and deployment of AL-based systems in healthcare [26, 82, 109, 169, 221, 253, 299, 325]. While this is a turn away from a technology-centred approach and a turn towards a human-centred and sociotechnical approach to AI innovation in healthcare, the orientation is not entirely new. Berg and colleagues [31, 33], who studied the introduction of EHRs in the 1990s, have been influential with their conceptualisation of healthcare work as "messy" and "ad hoc" in nature and as an interrelated assembly of humans and things. They argued that attempts to structure this work through the formal, standardised, and rational nature of IT systems is challenging and that optimal utilisation of health IT applications is *"dependent on the meticulous interrelation of the system's functioning with the skilled and pragmatically oriented work of health care professionals"* [31, 243]. They proposed to undertake a sociotechnical approach [243] to systems development projects and to employ participatory design for early and continuous facilitation of user involvement, as discussed above.

Unique sociotechnical challenges of clinician-facing ML-based systems. The more recent sociotechnical turn and the call for participatory design in medical ML innovation is therefore rather a re-turn and an effort to alert fellow researchers, designers, and practitioners not to reproduce the age-old mistakes. However, our systematic review and analysis of the emergent challenges demonstrate that sociotechnical issues with healthcare IT are not only reproduced but also exacerbated by the introduction of machine learning and large-scale healthcare data. The added technical elements like healthcare data, ML models, and ML-based user interfaces increase the complexity of successfully designing, developing, and deploying clinician-facing IT systems.

The five sociotechnical challenges do have similarities with well-known HCI issues like difficulties with clinical workflow integration, not addressing the needs of clinical end-users, or inability to demonstrate improved clinical outcomes. However, they also signify the uniqueness of how ML-based systems are inherently more challenging to realise as part of real-world clinical practice than traditional non-ML-based systems. For example, the dependencies between good quality training data, the ML model and the perceived clinical usefulness, or the dependencies between the ML-based user interface and the achievement of interpretability and trust among clinical end-users. Other examples of what is uniquely at play include the increased need for end-user training and adhering to existing

attitudes and feelings about machine learning and automation. Lastly, the interdependencies highlight how real-world deployment and commercialisation hinge on ethical concerns, the interaction with the medical professional and academic arenas, and the unique legislation practices for approving ML-based systems in healthcare.

12.4.3.1 *Implication for HCI: Need for iterative co-configuration of sociotechnical system*

HCI has contributed to mitigating some of the sociotechnical challenges. To a large extent, research on human-AI interaction has revolved around the problem of the interface between the human end-user and the AI- or ML-based system. Research provided guidelines for increasing acceptance of AI-infused systems by improving the interpretability of ML output through explainable and interactive interfaces (see e.g. [6, 79, 191, 213, 336]. Lim et al. [213], for example, recommended generating reasoning explanations to novice users for improvement of understanding and trust in the system. Amershi et al. [8] provided guidelines for designing effective human-AI interaction. Other works have developed models and principles for interactivity with explanations to improve users' comprehension by allowing them to explore an ML algorithm behaviour through visualisations or interactive interfaces [79, 106, 195]. Research also presented concrete frameworks and lessons that can be used to address the uncertainties and help focus other early-stage collaborative activities. Yang et al. [355] propose and demonstrate the usefulness of a conceptual framework for discovering and assessing potential design challenges. They suggest a two-by-two matrix that uses two attributes of AI projects that are central to the struggles of human-AI interaction design: capability uncertainty and output complexity. Similarly, Mohseni et al. [235] offer high-level guidelines for multidisciplinary teams on building explainable AI-based (XAI) systems. Their framework links the design goals of XAI with ready-to-use evaluation methods.

While this work provides significant help in tackling some of the unique challenges, the sociotechnical lens has several implications for HCI and the interdisciplinary innovation process of clinician-facing ML-based systems. There is a need for extending the design space from focusing on the ML-based interface and the interactions to focusing on the "interrelation" [31, 32] between the ML-based system and the corresponding workflows, organisation, healthcare institution, and political arenas. There are several opportunities in attending to the interrelations [31, 32] or the "sociotechnical configurations" [46, 319], wich we discuss in the following. First, there is a need to approach the innovation process as one of organisational change and as a matter of "growing" [107] and "configuring" working relations [46] between data, ML models, user interfaces, users, workflows, and so forth. Second, there is a need for deploying mock-ups, prototypes and early versions of ML-based systems close to or within real-world clinical work practices. Third, there is a need for merging activities

of design, development, and deployment in an iterative and cyclical process.

Opportunity #3: Approach medical ML-based innovation as a process of sociotechnical co-configuration. To mitigate the unique challenges of making ML-based system work in everyday clinical practices, we suggest approaching the overall innovation process as a process of co-configuration. It can be crucial for successful innovation to recognise that the design, development, and deployment of ML-based systems in clinical environments does not happen in distinct and separate phases but during the continuous arrangement and adjustment of working sociotechnical relations. The metaphor of "growing" [107] has been proposed as an attempt to capture the somewhat organic ways in which a new IT system needs to be adapted, cultivated, and reshaped jointly with the existing environment. An example from the reviewed articles is the work on Sepsis Watch. Such continuous and meaningful collaboration during design, development, and deployment resulted in an effective and trustful ML-based system [290, 301, 302]. This extended view of the innovation process, as one of configuration [311], or what we would like to call co-configuration, implies that success with ML in medical contexts is a matter of continuous efforts of evolving the components of the technology (e.g. training data, ML models, user interface) alongside the social dimensions of the environment (e.g. existing clinical practices, workflows, and organisation).

Opportunity #4: Near-live and real-world experimentation is necessary for innovation of clinician-facing ML-based systems. A second important strategy to mitigate the sociotechnical challenges is the commitment to introducing prototypes, paper mock-ups, and early versions of ML-based systems close to or within real-world clinical contexts. This systematic review evinces that critical insight to support an overall process of ML-based innovation is discovered only by the qualitative engagement with some form of test or evaluation with healthcare professionals. This can happen either in a lab, a clinical setting, or as part of everyday clinical work. With this review, we find that lab-based experiments and evaluations circumvent several difficulties with fully anticipating the impact of a complete ML-based system. This means that early testing with mock-ups or ML-based prototypes in lab environments provides critical, although speculative, insights for further design, development, and deployment. However, it is the near-live (see e.g. [221]) and actual deployments (see e.g. [26, 109]) that provide real-world evidence and opportunities for engaging in re-design and appropriation of the ML-based system and corresponding workflows. This proposal of striving for early in-the-wild evaluation, and ideally deployment in authentic healthcare settings, has been collectively put forward before [115]. However, the unique challenges with machine learning, e.g., the unanticipated outcomes of ML models and their impact on clinical workflows, require special attention to real-world experimentation.

Recent HCI work on clinician-facing ML-based systems follows a similar line of action. Studies of specific clinical decision-making

processes propose several concrete recommendations which can help navigate the uncertainties of AI/ML design and real-world deployment [135, 169, 356]. They suggest that intelligent decision support technologies should be tailored for a specific time, place, and decision context rather than pursuing a one-size-fits-all approach. Similarly, some of the included articles described such adjustments as an opportunity to engage clinical end-users and induce their trust in the system [67, 221, 286]. Researchers also highlight the need for minimising the extra effort incurred by the operationalisation of the ML output. Yang et al. [356] described it as unremarkableness, i.e., being situated naturally in an existing decision-making routine and only noticing when it might add value to the decision. Similarly, Jacobs et al. [169] described how the ML output and clinical action needed to be connected in ways that support workflows, which often involved additional healthcare providers.

Opportunity #5: Merge activities of design, development, and evaluation in an iterative and cyclical process. Striving for generating real-world evidence early on is, we argue, crucial for succeeding in tackling the inevitable challenges that emerge when transitioning laboratory ML models and systems into clinical settings. A final implication, derived from the identified sociotechnical challenges and related literature, is the need for merging activities of design, development, and evaluation. In their recent work, Elish and Watkins [109] raise a similar argument based on their participation in getting a deep learning model for sepsis detection to work in the wild. Undertaking sociotechnical interventions, they emphasise, is necessary to counter the risk of ML-based medical systems remaining potential solutions. They propose the concept of "repair work" to emphasise that innovation occurs *"throughout the implementation process and not just in the research or design phase"*. With their proposal, they increase attention to the skills, background, and invisible work required to make the ML-based system work for clinical practice. Earlier work has raised comparable advice by proposing a "radical refiguring of the relations of design and use" and by recognising the extent to which design activities must continue after the system has been deployed and put into use [319]. Other research has similarly proposed processes of "co-realisation" [144], "bricolage" [63], and "bootstrapping" [143]. Collectively, this research argues that innovation of disruptive technologies only happens by committing to design-in-use and real-world interventionist experiments. Inspired by this work, and taking into account the unique problems encountered in the included literature, we propose to carefully consider concrete ways of merging activities of design, development, and evaluation in an iterative and cyclical process when engaging in the innovation of clinician-facing ML-based systems. It is, nonetheless, critical that this methodological rethinking considers ethics (see e.g. [225]) and regulatory oversight to manage patient risk and to ensure final approval of what essentially will be classified as a medical device [238]. This can be achieved by committing to robust clinical evaluations that aim at the high quality of care and attractive patient outcomes [184, 329].

12.5 LIMITATIONS

The aim of this qualitative systematic review was to unpack studies on the real-world implications of concrete ML-based medical systems. Taking an HCI perspective on medical ML innovation, we excluded articles that reported solely quantitative data. While this criterion aligns with the overall goal of this paper, the descriptions of the technical aspects and medical specialities may be incomplete, overlooking systems that were evaluated only through quantitative studies. Moreover, due to the ambiguous nomenclature used in computer science and Health to describe ML development, we decided to search for relevant publications using broad queries, ensuring we did not miss any relevant publications. The queries yielded a considerable number of articles (9,672); hence, we restrained our search to three databases where one was covering Health an two were covering computer science. We acknowledge that this decision may have resulted in some publications missing from this review.

12.6 CONCLUSIONS

Recent years have seen a resurgence of interest in ML-based systems for clinical practice. Lab-based studies have provided promising results and suggest that ML outperform statistical methods and is capable of supporting the work of healthcare professionals and improve clinical outcomes. However, clinician-facing ML-based systems are particularly challenging to realise in clinical practice and, despite the favourable outlook, ML-based systems have not been widely adopted. To support researchers and practitioners from the HCI, ML, and Health domains in ML innovation, we systematically and qualitatively analysed articles that investigated the real-world implications of concrete ML-based medical systems. The compilation of 25 articles provided a comprehensive overview and deep insights into the challenges and opportunities for design, development, and deployment of ML in healthcare settings.

Through the reviewed literature, we identified key difficulties with medical ML innovation. First, an interdisciplinary collaboration among HCI, ML, and Health is particularly challenging and constituted by: technical choices, the intended role of ML, the activities and techniques applied, and the ways in which clinical end-users and other relevant stakeholders are engaged in the innovation process. Based on grounded theory analysis, we developed a semantically rich conceptual framework that, by our suggestion, can be instrumental for medical ML innovation processes. We conclude that shared terminology and striving for mutual understanding among project participants are pivotal to the realisation of medical ML innovation. Second, there are certain sociotechnical interdependencies that, if not addressed, can hinder the successful clinical adoption of ML-based systems. Mitigating these complexities require new modes of interdisciplinary collaboration. Opportunities for successful

ML-based innovation, we suggest, can happen through iterative co-configuration and near-live and real-world experimentation. We call on the HCI community to take the lead in the development of novel, yet much-needed participatory design principles, methods, and techniques to contribute to going the last mile of realising ML in healthcare settings.

ACKNOWLEDGEMENTS

The authors want to thank information specialist Julie Kiersgaard from the Royal Danish Library for her valuable assistance in searching for relevant literature.

13

PAPER II: GROUND TRUTH OR DARE

TITLE

Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI

AUTHORS

Hubert D. Zając, Natalia R. Avlona, Finn Kensing, Tariq O. Andersen, Irina Shklovski

DOI

https://doi.org/10.1145/3600211.3604766

VENUE

Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society

PUBLISHED

29 August 2023

ABSTRACT

One of the core goals of responsible AI development is ensuring high-quality training datasets. Many researchers have pointed to the importance of the annotation step in the creation of high-quality data, but less attention has been paid to the work that enables data annotation. We define this work as the design of ground truth schema and explore the challenges involved in the creation of datasets in the medical domain even before any annotations are made. Based on extensive work in three health-tech organisations, we describe five external and internal factors that condition medical dataset creation processes. Three external factors include regulatory constraints, the context of creation and use, and commercial and operational pressures. These factors condition medical data collection and shape the ground truth schema design. Two internal factors include epistemic differences and limits of labelling. These directly shape the design of the ground truth schema. Discussions of what constitutes high-quality data need to pay attention to the factors that shape and constrain what is possible to be created, to ensure responsible AI design.

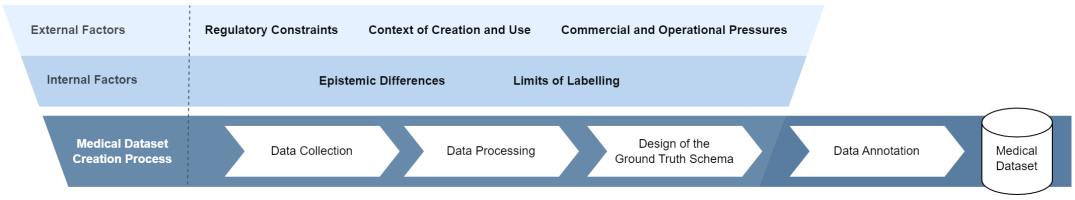


Figure 14: A simplified medical dataset creation process expanded with the design of ground truth schema and factors conditioning the pre-annotation stages.

13.1 INTRODUCTION

Advances in applications of Artificial Intelligence (AI) in the medical domain promise to improve efficiency, promote accuracy and bring cost savings across many areas of medical subspecialty, yet there are also many concerns about ethics and responsibility in the deployment of these technologies [96]. The idea of responsible AI has been extensively discussed in the literature and received much attention from both commercial entities and regulatory bodies [92, 234]. There is considerable agreement that high-quality training data is key to the development of responsible AI systems [156]. Yet research shows that the creation of high-quality data also tends to be an undervalued step in the development of machine learning systems [289, 293].

The process of dataset creation is typically broken down into three steps - data collection, data pre-processing and cleaning, and finally, data annotation[6, 245]. This is especially so in the medical domain where high-quality training data is obtained through a range of annotation practices such as data quality enhancement [75], generating labels using Natural Language Processing models [139], deriving image labels from medical documentation [170], and following labelling guidelines and principles focusing on fairness and inclusion [202, 295]. This paper investigates the factors that affect the creation of high-quality medical datasets demonstrating that the preparatory work involved in the design of ground truth schema used in data annotation is an important preceding step that tends to be overlooked in the literature. Following the work of Mueller and colleagues [240], we define the ground truth schema as a collection of relational labels and metrics, as well as their definitions and examples that are used during data labelling.

Recent research on the creation of training datasets [125] has discussed annotation activities as a matter of power relations in projects crowdsourced in the Global South [228, 229, 232], the social design of labelled data by domain experts [240], and annotation process recommendations [119]. While understanding data annotation is important, data design work begins before the first data points are labelled. Data is always designed and constructed through situated and emergent processes [114, 240] as domain experts, data scientists, other stakeholders, and diverse political interests imprint their values on the data. However, little is known about the preparatory work necessary to produce high-quality data [164]. Accounts of decisions that shaped

the datasets are rarely documented and get dismissed as soon as the data creation work concludes [293],thus become impossible to inspect in the future [240, 250, 310].

In this article, we consider *what factors affect the design of medical datasets prior to data annotation.* We ground our findings in ethnographic research conducted across three organisations developing medical AI for (I) screening chest x-rays, (II) supporting the diagnosis of lung and pancreatic diseases (III) automating patients-to-clinical trials matchmaking. We explore the decisions made by medical professionals, data scientists, designers, and other relevant stakeholders in their quest to create medical AI datasets in highly constrained environments. Our data include approximately 50 hours of observations, interviews with 46 medical professionals, data scientists, and designers, as well as observation notes, email communication, reports, and artefacts. We followed a grounded theory approach [73]that led us to identify and define factors that influence the design of the ground truth schema that underpins the production of high-quality training data.

Our contribution is twofold. First, we identify five factors, three external and two internal, that influence medical dataset creation by affecting data collection, ground truth schema design, and data annotation stages (see Figure 14). The external factors condition the medical dataset creation processes by determining the data collection and shaping the possibilities for the design of ground truth schemas:

- Regulatory Constraints
- Context of Creation and Use
- Commercial and Operational Pressures

The internal factors define the negotiations between the medical and technical domains:

- Epistemic Differences
- Limits of Labelling

Second, we show how these factors affect the final shape and quality of the resulting medical datasets. While we define each factor separately for analytical purposes, the factors are interrelated and affect each other, structuring the limits of responsible data creation approaches. We argue that these factors condition the stages that precede data labelling and mediate the design of what is aspired to be responsible AI.

13.2 RELATED WORK

While the idea of responsible AI has received much attention from both commercial entities and regulatory bodies, concerns about the quality of data and the challenges in the creation of quality data are

increasingly in focus. The now-emerging guidelines list several data-related challenges as key obstacles that hinder the path towards responsible AI: skewed data (issues that originate during data collection), tainted data (issues that stem from labelling e.g. hidden stratification [251]), or limited features (an inadequate number of features represented in data) [24]. There is broad agreement that dataset creation processes deserve greater attention, despite scholars repeatedly pointing to a strong bias against data work [91, 289, 293].

13.2.1 *How datasets are created and annotated*

In computer science, dataset creation is often seen as an activity constituting a step in the larger development processes of ML-based systems [6, 76, 152, 245, 335]. However, scholars have also discussed the dataset creation process on its own merits. For example, Hutchinson drew parallels between software development and dataset creation practices by sharing conceptual stages like requirement analysis, design, implementation, testing, and maintenance[164]. Similarly, increased focus can be observed in the medical area, where researchers describe in greater detail the creation of publicly available medical datasets [64, 90, 167, 173, 227, 339]. Typically, dataset creation is described as a process that spans all activities related to work on medical data, collected under the umbrella of data collection, data cleaning and processing, and data annotation.

Data annotation is one of the most researched aspects of dataset creation. Data annotation or labelling usually happens as part of the curation or preparation step of larger data science projects, following data acquisition and cleaning, and preceding feature engineering [6]. These activities are usually iterative and highly collaborative. Linguistic scholars and Natural Language Processing researchers [119, 161, 332] offer guidance on how to carry out data labelling. They distinguish three focal points: the creation and improvement of an annotation guide [119], schema [332], or manual [161]; the labelling performed by trained annotators; and the adjudication of the annotated data.

In this paper, we use the terms data labelling and data annotation interchangeably and understand them as the action of assigning and adjudicating predefined labels to concrete data points. When considering this step alone, there is a multitude of decisions that need to be taken to complete it. Scholars have pointed to data annotation activities as a site of political struggle, challenges to the labour conditions, as well as the stage in dataset creation that can result in adverse downstream outcomes for trained models. For example, Schumann et al. [295] and Hanley et al. [142] demonstrate how the design of categories (or labels) can reinforce harmful stereotypes and exclude underrepresented groups of people. Badly annotated data can reduce the performance of AI models [75, 139, 170, 234, 276] and perpetuate exclusion and inequality [202, 295].

In the medical domain, data annotation challenges can be compounded by the requirements for specialised knowledge and training. Despite initiatives like the Unified Medical System [214], the clinical meaning of labels can be unclear [250], and medical knowledge remains difficult to capture for computer use. Li and colleagues [211] explored the inter- and intra-rater agreement between six radiologists of different experience levels when labelling chest x-rays. In some cases, even the experienced radiologists reached only a moderate level of agreement with themselves [224]. This could occur due to not following the best medical practices when labelling data, due to resource constraints [289] or because of the disconnect between the practices of labelling and the actual usage of medical data in regular practice [250].

What much of this research points to is the fact that labelling and annotation as practices are heavily reliant on the creation of annotation guides and schemas [119]. Yet, despite the growing interest in the creation of datasets, current discussions tend to omit and overlook the pre-labelling activities and their potential impact on the quality of the resulting training data [359].

13.2.2 *The design of the ground truth schema*

Many scholars investigated the dynamic and situated work of domain experts, data scientists, designers, and other stakeholders engaged with data [150, 239, 289, 298]. For example, Muller and colleagues investigated how domain experts label data, highlighting that the ground truth contained in datasets is a human contribution resulting from improvised and iterative adjustments to principled design processes [240]. Discussing the design of ground truth schema implies that ground truth captured in medical AI datasets is not an objective representation of reality but is a result of a situated design process [54]. In other words, data is never raw [129], instead, all data is actively constructed [15, 230, 264]. Feinberg emphasises the importance of recognising the subjectivity involved in dataset creation and the need to consider the potential biases and limitations inherent in choices that stem from the social and organisational context in which data is produced [114].

Researchers who investigate AI datasets suggest that access to all of the "subtle design decisions", made during the dataset creation, is vital to ensuring a high-quality labelling process [112, 250] and thus high-quality datasets. However, documenting design decisions in data science work is not common [264, 287, 364]. To address this gap, researchers developed a range of documentation frameworks to support the accountability, use, and maintenance of complex datasets [11, 231]. These frameworks range from general purpose and qualitative - Datasheets for Datasets [123], NLP-focused - Data Statements [29], quantitative - Dataset Nutrition Label [154], to fairness focused - data briefs [112] and accountability [164]. Some of these tools [112, 123, 164] include a query for the origin of the labels, but most do not

pay much attention to the pre-labelling activities involved in annotation schema creation.

While the existing scholarship has problematised the stage of the data labelling and the power relations and conditions affecting the data annotation work [240], little is known about the stages preceding the data labelling. Particularly, how these stages influence the final shape of medical datasets. We explore the collaborative and situated work of medical professionals, data scientists, and designers that takes place before the labelling stage, within the design stage proposed by Hutchinson et al. [164] or the preparatory work proposed by Fort [119].

13.3 METHODOLOGY

We investigated three organisations in the Global North developing medical AI-based systems that engaged in the medical dataset creation processes. We focused on the work conducted before the data annotation task by participants described in Table 12.

13.3.1 *Research context and data collection*

13.3.1.1 *ORG I*

was an interdisciplinary collaboration between academia, business, and the public healthcare sector, aiming to create AI-based chest x-ray prioritisation software for global use. The project's first step was designing the ground truth schema for labelling chest x-rays, which is the process investigated in this study.

Our engagement in ORG I spanned May 2021 to Feb 2023. During that time, we conducted participatory observations of the design process of the ground truth schema. The working group developing the system was based in a Northern European country (Table 12.1). Additionally, a feedback group comprising medical professionals from the Northern European country and an East African country provided feedback on the schema (Table 12.2). We participated in fifteen working group meetings ranging from 26 minutes to 2 hours and 12 minutes in length. Additionally, we conducted twelve interviews and observed external medical professionals evaluating and providing feedback on the intermediate results of the design work. Additional material included observation notes, meeting summaries from other participants, a work progress report, email communication, and produced artefacts - a labelling guide and the ground truth schema.

13.3.1.2 *ORG II*

was a large tech company in Western Europe with part of the business involved in the development of complex medical devices. We primarily engaged with sections of the company that focused on the development of AI-based diagnostic tools and systems for oncological radiology.

ORG I	POSITION	EXP.
P1	Radiologist	Junior
P2	ML Engineer	Senior
P3	ML Engineer	Senior
P4	Computer Scientist	Senior
P5	Data Scientist	Senior
P6	Radiologist	Senior
P7	Radiologist	Senior
P8	HCI Researcher	Junior
P9	HCI Researcher	Senior

ORG I	POSITION	EXP.
P10	Radiologist	Senior
P11	Radiologist	Junior
P12	Radiologist	Junior
P13	Radiologist	Mid
P14	Radiologist	Senior
P15	Radiologist	Senior
P16	Radiologist	Senior
P17	Radiologist	Senior
P18	Radiologist	Senior
P19	Radiologist	Senior
P20	Radiologist	Senior
P21	Physician	Junior

ORG II	POSITION	EXP.
P21	Data scientist	Senior
P22	Product Owner	Mid
P23	Strategic Designer	Senior
P24	Data scientist	Mid
P25	Usability Designer	Senior
P26	Data scientist	Senior
P27	Data scientist	Senior
P28	Data Designer	Mid
P29	Interaction Designer	Senior
P30	Data scientist	Senior
P31	Data Designer	Senior
P32	HCI Researcher	Mid
P33	Data Designer	Senior

ORG III	POSITION	EXP.
P34	Product owner	Mid
P35	Software Engineer	Junior
P36	Software Engineer	Mid
P37	Software Engineer	Mid
P38	Data Scientist	Mid
P39	Data Scientist	Senior
P40	UX Designer	Senior
P41	Software Developer	Mid
P42	Medical Operations	Senior
P43	Quality Assurance	Senior
P44	UX Designer	Mid
P45	Neurobiologist	Senior
P46	Product Owner	Mid

Table 12: List of participants, their simplified positions, and experience levels. Respectively in ORG I (working group), ORG I (feedback group, participants 10-14 were located in the northern European country, and participants 15-21 were located in the East African country), ORG II, and ORG III.

Our work with ORG II was split into a preliminary exploratory period online from February to May 2022 and in situ participant observations and semi-structured interviews conducted in June 2022 in a Western European country. Due to the size of the organisation, we employed snowball sampling. In ORG II, we conducted thirteen semi-structured interviews with experts (Table 12.3), with an average duration of 65 minutes.

13.3.1.3 *ORG III*

was a mid-size start-up in Western Europe that aimed at developing an AI-based platform for matching patients with advanced clinical trials for new drug and experimental procedure development. The company primarily dealt with two data sources. First, they collected data from medical practitioners and their patients. Second, they collected data from public registries in the EU and US and pharmaceutical companies about clinical trial requirements or experimental treatments. Their goal was to match the patients with unmet medical needs and their physicians with the requirements of BioPharma companies that need to enhance drug development and recruit participants for clinical trials.

Our engagement with ORG III spanned February to May 2022. The preliminary period involved online semi-formal meetings and interviews from February to April 2022. In-situ ethnographic research was conducted during May and June 2022 at the headquarters of ORG III in Western Europe. We conducted participant observation by joining the daily stand-up sessions of the engineering department and shadowing the workflow of the AI team experts leading the data labelling process for the match-making platform. In total, we interviewed 13 participants (Table 12.4).

13.3.2 *Data analysis*

The main focus of our analysis was to identify factors affecting medical dataset creation. We analysed decisions made during the design work, tensions and misunderstandings that needed to be reconciled, looking both outside and within the organisations where the design work took place. We explicitly decided to explore the wider socioeconomic factors that condition the medical dataset creation and influence the final AI-based systems even before the first label is annotated.

Data analysis relied on techniques of grounded theory and situational analysis [73, 81]. First, we conducted line-to-line open coding, coming up with 850 initial codes. We then reflexively proceeded to thematic coding, in an iterative manner, discussing the themes and patterns emerging in our three sites of ethnographic inquiry. During this step, we designed visual maps to lay out the human, technological, and discursive dynamics of the organisations under study [81]. Second, we conducted axial coding to reflexively group the available themes into dimensions. Finally, we assessed these dimensions

against the codes and situational maps, converging on the five final factors (regulatory constraints, context of creation and use, commercial and operational pressures, epistemic differences, and limits of labelling).

13.3.3 *Positionality statement*

Our qualitative data was obtained from three health-tech organisations in the Global North. The analysis was shaped by the following standpoints. First, we differentiated our roles in studying the three organisations. Researchers in ORG I had the dual position of the expert who on the one hand designed the labelling software, whilst they conducted participant observation and semi-constructed interviews in order to study the process of the ground truth schema design. Researchers working with ORG II and ORG III employed ethnographic methods as a research approach without having a prior engagement with the organisations. Second, we are researchers currently working for Northern European institutions. Third, we have mixed epistemic backgrounds in computer science and law and policy. Finally, we emphasise the situatedness of our research, which focuses on the development of medical AI at the specific loci of our studied organisations. We acknowledge that the factors we identify as defining the medical dataset creation bear the geographical and epistemic limitations of the Northern European context. On this note, we acknowledge that the divide between Global North and Global South we make below has been problematised by scholars in human geography and decolonial studies as a limiting one, reinforcing stereotypes and reducing the polyphony of southern standpoints [160, 333]. For this reason, we use this divide in this paper to (I) acknowledge the limitations of our standpoints in a northern institution and the privilege of our funded projects; (II) tackle assumptions about data universalism [233] by showing the particularities of the northern context in medical datasets creation and their effect on the intended use of such data in different contexts.

13.4 FINDINGS: FIVE FACTORS THAT INFLUENCE MEDICAL DATASET CREATION

The datasets used for medical AI benefit from the impression that they are a result of an age-old medical practice that is seamlessly transitioning to the digital age, unaffected by external influences, and focused on the pursuit of medical excellence. However, the reality is often different. Our ethnographic data suggest that even before medical professionals have had the chance to annotate or make their first label, many critical design decisions have been made, which frame the labelling space, thus limiting the extent to which medical professionals can use their expertise.

Our analysis challenged our initial understanding of the dataset creation process drawn from the literature. Our data made clear that

the preparatory work should be conceptualised as a crucial stage in dataset creation taking place before data labelling because it defines what becomes captured as ground truth within a training dataset. This is the step where the ground truth schema is designed, which, when applied to an unlabelled dataset through expert annotation, embeds the intended ground truth within it.

We identified five factors that influenced the creation of medical datasets in the organisations we studied. Three of these factors were external to the activities directly involved in pre-labelling activities. External factors defined and delineated the limits and possibilities for labelling activities. Two internal factors on the other hand affected the negotiations around what needed to be labelled and how the labelling was to proceed through the design of the schema. Below we describe each factor and demonstrate how they affected the final shape of the medical datasets focusing on the data collection and ground truth schema design stages.

It is important to note that the organisations and processes examined in this paper were largely driven by data scientists as the owners of the dataset creation process, with representatives of other domains contributing to the dataset creation activities. As a result, data science as an epistemology dominated the design work by primarily embedding data scientists' perspectives, inadvertently compromising other domain-based practices and understandings. As datasets in our research were created for the purpose of AI development, the power distribution was uneven, leaving little room for misconceptions from data scientists to be challenged and addressed.

13.4.1 *External factors: defining the ground truth schema design space*

Despite the best intentions of the experts engaged in the medical dataset creation process, many of their decisions and actions were structured by different external factors. We identified three such factors - **Regulatory Constraints, Context of Creation and Use**, and **Commercial and Operational Pressures** - that shaped the space of medical dataset creation and thus influenced the final shape of the datasets themselves even before the labelling could begin (Table 13). Each factor consists of several distinct features. We describe these below in detail.

13.4.1.1 *Regulatory constraints*

The medical data space is highly controlled through a variety of local, national, and international regulatory constraints. This was particularly challenging for the data collection step of the process. We observed two areas where compliance with regulatory standards affected the creation of medical data: **the extent of the collected data** and **the predetermination of purpose**. Experts in all of the organisations we studied were concerned about compliance with diverse standards that intersected with their work on medical dataset creation. These standards originated from European binding legislative acts, in-

REGULATORY CONSTRAINTS
Extent of Collected Data
Predetermination of Purpose
CONTEXT OF CREATION AND USE
Geographic context of use
Demographic context of production
Linguistic context
COMMERCIAL AND OPERATIONAL PRESSURES
Business model and organisation scalability
Competition and health tech market
Intended future use within the healthcare type

Table 13: External factors and their dimensions.

ternational standard organisations, or industry standards. GDPR, the main legal standard for data protection in the European Union, was the most prominent example of a binding legislative act, regulating the conditions under which personal data is collected and processed. The industry and international organisations imposed, among others, ISO 2700013001, HIPAA, and Good Medical or Good Manufacturing Practices. In ORG III, a data scientist (P39) listed 21 unique regulations they felt they needed to consider. As a larger and more mature organisation, ORG II also had internal ethics boards, which at times imposed even stricter interpretations. However, these standards and limits legitimised the data collection and processing activities.

Constraints on data collection. While experts in all organisations were striving to create what they saw as high-quality data, complying with relevant regulatory standards required concessions from all participants. For data scientists, the regulatory constraints delimited what data was available for collection, at times inadvertently introducing bias in different ways. For example, P26, a data scientist from ORG II, explained: *"what is the data that we are allowed to use, especially if you look at ... bias ... people will want to look at bias and, and see if ... their product was fair to all, some demographics, and [we are] just not able to use the data because of privacy issues or GDPR"*. Similarly, in ORG II, the contractual agreement with a single local hospital, on the one hand, provided a controlled supply of high-quality data, on the other hand, reduced data representativeness: *"we have a strong relationship with them. How do you expect that the data is not going to be biased right?"* (P24). While ORG II was able to create highly detailed and structured training data for their models, this data was clearly not representative of populations that would eventually encounter the resulting technologies.

Limitations imposed on data collection could compromise the resulting datasets in ways that created challenges for subsequent data creation steps. For example, participants of ORG I could collect only chest x-rays and their linked radiological reports. Privacy concerns here also resulted in the loss of the chronological links between the images during data collection. This selection significantly diverged from the usual assortment of data available to radiologists in clinical practice, introducing challenges at the later stages of medical dataset creation, such as schema creation and annotation.

Regulatory standards and contractual agreements determined the purpose and context of use. Data protection regulations have recently focused intently on the purpose of use as one area of emphasis, tied to notions of data minimisation and data subject notification. Companies in our research had to negotiate the legal basis for their data collection with contracted data providers such as hospitals. For example, GDPR and contractual agreements with a local hospital bounded ORG II to use the collected data within the predefined purpose and context. Deviations from the initially stated purposes and context of use required new agreements that could be obtained only through significant time and resource investments. As a product owner (P22) explained the process of collecting data from the local hospital, *"maybe the new study that we want to do has a slightly different scope and it's not covered by the original contract, then we need to make a new contract"*. ORG I encountered a similar predicament where the data collection phase was negotiated based on what the data scientists believed to be a necessary and sufficient dataset given the available resources and legal constraints of local regulations. By the time domain experts explained that the dataset was lacking important data dimensions, it was too late.

13.4.1.2 *Context of creation and use*

The context of production and the context of use influenced the creation of medical datasets. In our studies, each medical dataset was created for a specific intended use that was embedded in the collected medical data, e.g., clinical trial repositories, hospitals, and patients. These sources cover specific geographical populations, which has consequences for the final medical dataset. We identified three dimensions where that influence was prevalent: **the geographic context of use, the demographic context of production**, and **the linguistic context** (Table 13).

The geographic context of use affected the selection of labels. While medicine strives to deliver replicable results that generalise across populations, the ground truth schemas are designed to serve specific needs in specific contexts. Some of them are defined by the intended use of the future AI-based systems in the geographic context, in which they are going to be used. In ORG I, the project group designed the first version of the ground truth schema based on local data from a Northern European country. As a result, the first version of the schema captured the locally prevalent conditions well

but missed conditions relevant within the countries of intended use, which were almost never encountered locally. To account for that, direct and indirect input from medical professionals from the East African country was collected and incorporated into the schema during joint design work, as seen in this exchange between a radiologist and an ML engineer.
"So if you wanted that in the hierarchy, it could be there." (P1)
Is it aortic unfolding? Because I clearly remember this sentence from [the East African country] reports, "aortic unfolding due to chronic hypertension" (P2). Yet despite having a broader ground truth schema, the same project also struggled to ensure enough examples of common medical conditions across expected countries of use available for annotation, since the data was originally only collected from one country.

The demographic context affected representativeness concerns In both ORG I and ORG II, data in medical datasets were collected from a single country, which had several consequences. For example in ORG II, the data was predominantly collected from a single local hospital, where ORG II had a contractual agreement. Not only was this problematic due to a more homogeneous patient population, but the collected medical imaging data originated on machines from the same producer. This created many concerns since imaging machines from different manufacturers often produce slightly different artefacts in their output. Yet the information about which machines were used to produce the images was rarely included in the resulting dataset.

Similarly, due to the characteristics of the population embedded in medical datasets, experts worried about how portable the resulting AI models would be. As a usability designer (P25) from ORG II noted, *"you can have all sorts of differences in patient demographics ... and you cannot just apply a model that you train on population A to population B"*. However, despite the designers' and data scientists' awareness, a senior radiologist from the East African country emphasised that *"in the [developing world][1] we are usually consumers, not producers of tech. We may find ourselves hitched to tech that doesn't serve our needs"* (P15). When evaluating the ground truth schema, the same medical professional elaborated, *"I've done this for 10 years since my graduation. I've never seen certain diseases like cystic fibrosis, but whenever I read the books, there's a lot of stuff about cystic fibrosis [prevalent in the Global North],"* which highlights the effect of local ground truth schemas on the transferability of the final AI-based systems.

Linguistic context and local understanding of medical terms challenged the application and transferability of the ground truth schemas. The design of ground truth schemas included naming the labels, defining and organising their relations, and providing examples. However, medical concepts are not always used in the same way across different countries. In ORG I when discussing the naming convention for a chest x-ray finding, one radiologist noted

1 edited to avoid pejorative language

"I know that it's not proper, but [in the Northern European country] they use 'infiltrat' as a synonym of consolidation ... I think the direct translation consolidation would be 'consolidering' but they don't use that, they use 'infiltrat'... I think maybe our infiltrate is broader" (P1). As a result, a presentation of infiltration by an AI-based system could be understood differently by medical professionals from different countries. To account for that, data scientists and medical professionals evaluated the ground truth schema against English translations. In ORG III, which operates globally, the data scientists and designers recounted a similar challenge of re-translating medical terms during the data annotation process. The limitations of the locality of medical terms prohibited the aspiration of designing a ground truth schema that can operate universally. As a UX designer (P40) remarked: *"there are also challenges around that because different cultures will refer to different diseases in different ways. It's global and we re-translate some of our stuff into different pages. We also have to consider localisation, how you turn this medical term into a layman term, but that's also relevant in like different countries as well."*

13.4.1.3 *Commercial and operational pressures*

The three organisations each had a different business model and exhibited different relations to the market and the public sector. This often determined the availability of the resources (human and material) allocated for dataset creation and affected the organisations' ability to collect data and design the ground truth schema. We identified three dimensions of commercial and operational pressures (Table 13): **business model and scalability of the organisation, the competition in the health tech market**, and **intended future use within the healthcare type**.

The business model and scalability of the organisation determined the amount of collected and labelled data. Every investigated organisation represented a different business model. ORG I intersected with the public sector, whilst ORG II and III were situated entirely in the private sector. The business models of the organisation determined the way in which data was collected. The business model of ORG III relied on providing free use of the AI-based platform to patients but also providing paid services to BioPharma by enrolling patients into clinical trials. To do that, ORG III collected data from the public clinical trial registries in the EU and US, as well as patient medical information. Such data collection was heavily dependent on the organisation's scalability, as well as the "fine" balance between the data requested by their BioPharma clients and the data that could have been collected. As a data scientist (P38) explained: "sometimes it's difficult to decide what kind of data you collect, right? Or what patients. (...) there's a balance between what's actually feasible to collect and what will give us the highest chance of getting as much data as possible. So those I think are tricky decisions." These conditions affected how much data was finally collected, hence, the ideal of representativeness of the created dataset was compromised.

In ORG I, the budget allocation for the data annotation process played a vital role in the amount of data possible to be labelled by medical professionals. Due to the high cost of labelling by experienced medical professionals, ORG I had to cap the maximum number of labelled images. This cap limited the number of distinct labels that could be annotated in the created dataset and remained statistically significant. "We have a limited budget for the test data that we can collect because we need several radiologists board-certified possibly to look at images" (P3). The limited resources defined the amount of data that was possible to be annotated, putting ORG I at a competitive disadvantage: "What the [competitors] do (...) there is no way we can reach what they do. They have 127 findings and they use a hundred plus radiologists to annotate, and they annotated 800,000 images each image by three radiologists. So the scale is completely different" (P2).

Market standards and industry competition affect the design of the ground truth schemas. Since all organisations under study operated in the health tech sector, the experts engaged in the processes of designing ground truth schemas had to both consider existing state-of-the-art solutions and methods, as well as address market competition. In ORG I, the choice of a specific machine learning model architecture was dictated by the industry standard. However, this choice had consequences for the label needs during the design of the ground truth schemas. At the same time, addressing market competition influenced the work on the ground truth schema design, as seen here, "so this is [a competitor's system] and this is their output. they ... split consolidation and nodules, which at this stage of the hierarchy we are not doing. And so I was wondering why we're not doing it" (P2). In this organisation competition directly influenced the design work.

Due to the large size of ORG II, the matter of competition fed to internal business processes whose results other experts relied on during the dataset creation, as explained by a product owner (P22), "it's a combination of ... alignment with the business priorities and that is also strongly driven by customer requests and customer demands. So that is actually very important ... try to find the alignment". Finally, market competition created time pressures that could structure and limit how data creation had to be organised: "if you want to validate something properly, it costs time. If you want to validate across domains, it costs time. And we are often in very competitive domains where being fast to market or, or fast at the FDA is also important. So there are some time trade-offs, need to be made there." (P27).

The intended use and type of healthcare system affected the content and the level of detail of the ground truth schemas. Visions of future intended use permeated the design work on the ground truth schemas. The imagined intended use of a future AI-based system factored into decisions about the validity of label choices. Imagined use did not fit in with current domain-specific practices and resulted in confusion and concerns during the design of the ground truth schema. Consider the following discussion between a medical professional and data scientists from ORG I about the implication of

different intended uses of the future system for the selection of labels. *We have two priorities, one is decision support. So it might be easy for you to see the mass, so that won't help you. But there's also the pre-screening - prioritisation. So that might be relevant to detect mass prematurely, right?* (P3)
So if you use it for like a warning, a prioritisation, it can be useful, but for detection... we can see a mass. It's not difficult to find (P1).

Medical AI-based systems in our organisations were designed to operate across the world within public or private healthcare systems. Yet medical systems in different countries operate differently based on public values, profit, incentives, and conventions. The design decisions during dataset creation are a product of all these components. The dependency on the healthcare type was well captured by a data scientist from ORG I when discussing the level of detail of the ground truth schema, *"if it was in the US where you actually pay, then from a business point of view, you really wanna find everything. First of all, you don't get sued, and secondly, you can make a lot of money by treating them. But here it's very different, right? Because it's a public system and you only treat things that are necessary, that need to be treated, right?"* (P4). These concerns manifested in debates about what could and needed to be annotated as expert annotators infused the values of their local system into data creation activities.

13.4.2 *Internal factors: designing the ground truth schema*

While external factors were key in shaping what data was collected and made available for annotation and highlighted the importance of local considerations and their implication for the resulting datasets, two internal factors drove debates, discussions, and disagreements that affected the ground truth schema and the resulting datasets. These were **Epistemic Differences** and **Limits of Labelling** (Table 14). The effort going into the creation of medical datasets as training data had two purposes that sometimes came into conflict. First, medical datasets were seen as a means of capturing the current state of medical knowledge and the tacit knowledge of medical professionals who focused on medical practice and clinical usefulness. Second, the same datasets served computer scientists as complex input data to solve problems through mathematical operations, where consistency and accuracy were in the spotlight. These two perspectives, while not opposing, often prioritised distinct qualities of the same datasets.

13.4.2.1 *Epistemic differences*

While in ORG II and ORG III, we engaged with relatively homogeneous teams within each company, in ORG I, our research process was focused on supporting the data creation process by working together with the data science and radiologist teams. As such, in ORG I, we were able to observe first-hand how teams with domain expertise often disagreed on what constituted legitimate knowledge as they discussed what was worth annotating and how things ought to

EPISTEMIC DIFFERENCES
Miscommunication between domains
Misapprehension of medical practice
Misapprehension of medical knowledge
LIMITS OF LABELLING
Domain expert buy-in
Onboarding to the labelling task
Labelling hardware and software
Similarity to the clinical practice

Table 14: Internal factors and their dimensions.

be annotated. We consider three sources of epistemic differences that affected the final design of the ground truth schemas (Table 14), communication challenges within the teams, misapprehension of medical practice, and misapprehension of medical knowledge. Within these dimensions, team members from different domains expressed diverging priorities, values, and understanding of concepts, which needed to be reassured and negotiated.

Communication challenges within teams. The three organisations involved stakeholders from different backgrounds, such as health, data science, and design. All of these brought their own traditions, meanings, and domain knowledge that needed to be shared, translated, and understood by other parties for worthwhile collaboration. It is no secret that interdisciplinary teams must spend time finding common ground before they can work together productively [59]. In our research, we observed how medical professionals, designers, and data scientists constantly translated and explained concepts from their respective domains to maintain a shared understanding. For example, at the beginning of the study in ORG I, medical professionals designed labels based on their, at times naive assumptions of machine learning capabilities, such as when they included two medical concepts under the same label, *"but couldn't that be, if you put nodule, mass in the same category, couldn't you just program it, later on, to say that if the thing that they have marked nodule/mass is over I think ... five millimetres or something, you call it a mass"* (P1), which was not possible given the collected data and was later clarified through a joint discussion. Similarly in ORG II medical professionals had to explain to data scientists that to detect some types of cancer it is necessary to look at more than just the organ in question, and that doctors need to use other information, such as the condition of bile ducts or the blood flow around the organ, affecting data collection and subsequent labelling set up.

Misapprehension of medical practice. Across the organisations the expectations for the quality of the datasets were closely aligned with concepts such as consistency or bias. This focus was clearly visible when discussing the goal of the labelling task in ORG 1. In the pur-

suit of consistent and unbiased data, data scientists initially framed labelling as a *"different task"* to clinical work: *"We need to know what's in the image and we need it without them being biased towards looking for only stasis"* (P6). As a result, the labelling task did not provide what was seen by the data scientists as "extraneous and potentially biasing" information, such as the background information of a patient. However, situating the labelling task further away from the medical practice affected the quality of the input medical professionals could provide, impairing the ability of medical professionals to use their knowledge. As one senior radiologist (P10) noted: *"Asking a radiologist to categorise something on a picture only without getting any information on the patient. Is like asking a surgeon to look at the scars on a patient and having him tell you what kind of surgery that patient had"*.

The pursuit of objective and unbiased labels isolated labelling from what data scientists saw as extraneous, potentially biasing information. Yet this transformed the work of the radiologists into a new task that was incompatible with medical practice. To deliver the expected results in this new unfamiliar process, radiologists attempted to reconstruct their medical practice by drawing from their tacit knowledge or, simply, guessing: *I have to create something about the patient myself, which is, [or] might not be true. And I then describe the picture from there...* (P10).

Misapprehension of medical knowledge. Specific data was needed to train AI models that provide clinically useful functionalities. However, due to the misapprehension of practice, the assumptions about what clinical knowledge was possible to extract from the clinical data provided were also at times flawed. As the schema went through iterative rounds of design, we observed how both sides struggled to understand why particular data was requested or why a particular request seemed to be difficult to fulfil. For example, in ORG I, radiologists were asked to assign one of three possible values as a patient's general state based solely on a single chest x-ray, so that relevant cases could be later prioritised using the resulting AI system. This task proved to be particularly problematic to radiologists who do not use such metrics in their daily practice, so they had to develop a range of new approaches to assign them, like *"I chose to interpret it from the view that it could be the worst situation"* (P12) or *"I think it was mostly a gut feeling"* (P11). In the end, the radiologists produced the kind of data that data scientists expected to see as labels. However, what these labels actually captured diverged from the original intention.

13.4.2.2 *Limits of labelling*

Finally, we turn to the mechanics of labelling itself that affected the final design of the ground truth schema. We observed schema design and testing in situ directly in ORG I, while in ORG II and ORG III, our data come from post-hoc interviews. We find that four features affected the final design of the ground truth schema (Table 14), domain expert buy-in, onboarding to the labelling task, clinical practice

familiarity, and labelling hardware and software. These dimensions manifested when evaluating the labelling processes. Unlike the *Epistemic Differences*, where data science was the defining domain, the *Limits of Labelling* emerged as medical professionals confronted the intermediate results of the epistemic negotiations discussed above. These limits altered what kind of data was collected and affected the quality of the labelling.

Domain expert buy-in. Our data showed that domain expert buy-in was crucial and required concessions on the type and amount of collected data. Some ML models require specific types of annotated data, such as *"what we're asking them is for each patient to go through 500 images and for each image to annotate [...] at pixel level"* (P21). Not only are such tasks typically outside of the scope of clinical practice but are also mentally challenging. For example, when P1 was asked to oversee the labelling process performed by external radiologists, they recalled: *"I think that he [a senior radiologist] opened the program, saw how difficult it was, and just closed it and just never had the energy to start it again"* (P1). Monetary compensation turned out to be a necessary but not sufficient strategy in ORG I for recruiting medical professionals with high expertise to annotate data.

Once the experts agreed to annotate data, **limited training for the labelling task reduced the chance for a "shared mindset".** Additional metrics were a relevant part of the ground truth schemas. These metrics usually included concepts not used in daily clinical practice. In ORG I, the medical professionals were supplied with written guidelines to boost common understanding and were briefly introduced to the labelling task. However, some annotators referred to the guidelines only when in doubt: *[the labelling software worked] right out of the box ... I didn't really read this part because it was not necessary* (P12). Not knowing the exact guidelines, medical professionals relied on an intuitive understanding of the metrics and labels, which often resulted in discrepancies between the annotators as they attributed different meanings.

Hardware configuration and user interface of the labelling software affected the quality of the annotations. These challenges were observed to a greater extent in ORG I, as to assess medical data like CT scans and x-rays, radiologists usually use diagnostic displays. Thus, when they annotate on a *"non-diagnostic screen, you miss details ... maybe small, smaller changes would be missed ... we don't annotate them because we cannot see them"* (P13). Similar comments were shared during the evaluation of the labelling software, medical professionals marked the location of findings using touchpads, which resulted in frustration and low precision.

Labelling software design could have influenced the final quality of the medical dataset to an even greater extent if not caught during the evaluation. Labelling medical data requires "[a] professional tool that could do the job in a very efficient way" (P21). However, the design of this software could have influenced radiologists in ORG I to overreport radiological findings per x-ray during an evaluation

period, *"...maybe it's an interface. Maybe they forgot the normal button was there because they only saw the [labels]"* (P1).

The overreporting was not solely caused by the labelling software. **Expectations and habits influenced what medical professionals noticed in medical data.** For example, a radiologist who reported on an evaluation of the ground truth schema in ORG I reported, *"I told my participants that there would be some normal, but they have not marked any of them normal or I can't find them"* (P1). This phenomenon was later explained by a senior radiologist who pointed to the expectation of labelling a dataset with findings and the fact that when the ratio of abnormal to normal cases is skewed, radiologists tend to overreport to remain on the safe side, *"that's [why] they thought they saw something that was not there"* (P6).

13.5 DISCUSSION

In the creation of high-quality training data, our research shows that the design of ground-truth schema is a crucial but often overlooked step. We highlight five factors that represent external and internal constraints that directly affect the quality of the resulting medical datasets. The external constraints condition the data collection process, affecting this way the design of the ground truth schema, while the internal constraints strongly affect the resulting ground truth schema and can lead to disagreements and debates among domain experts, predominantly data scientists and medical professionals.

13.5.1 *Conditioning the data collection*

Our findings demonstrate that the regulatory constraints, along with the geographical, demographic, and linguistic context of creation and intended use, and the organisations' scalability crucially affect the amount and type of data that was possible to be collected by the organisations we studied. In this sense, specific data quality metrics were already compromised since the first stage of the medical datasets creation. For example, in ORG I and II the geographical and demographic distribution of the collected data reflected not only how much data was possible to be collected by the contractual agreements in place but also manifested a lack of representativeness, given the regional and local source of data collection.

In ORG III, the aspirations for creating datasets of global coverage stumbled upon the linguistic contextuality of medical terms, which proved to become an issue during the ground truth schema design for the match-making platform. Similarly, in ORG I, the geographical, demographic, and linguistic context of the medical data collection shaped the type of the collected data, such as that when the experts came to decide on how to design the ground truth schema, dilemmas did not only concern the different understanding of the same medical terms across countries and continents but also possible omissions of local lung diseases. In this sense, the aspiration of designing "transfer-

able" ground truth schemas proved to be both dependent and limited by the standards that regulate the data collection and the context of its collection.

A further insight that emerged in our studies was that the business models and scalability of each organisation affected differently its capacity to collect data. For example, ORG I, being a small-size start-up, having however the public sector involved in its entity, had easier access to timely data (x-ray images of multiple years) from regional hospitals. Yet, the organisation's limited scalability defined the amount of data that was possible to be labelled by medical professionals. In Org III, a similarly small-size start-up, the data collection from both public registries and patients was shaped by the organisation's availability of resources. The constraints were imposed on the recruitment of data scientists designing the platform's ground truth schema and medical professionals who assisted the patients in submitting their medical information into an appropriate and structured format. On the other hand, in ORG II, due to its large size and scalability, the limitations of the data collection were shaped by market demands. This was reflected in the need to collect quality data, i.e., particularly structured, consistent, and contextual medical images from a controlled environment (the contracted local hospital). This push for one type of quality reduced another, in this case, the representativeness of the acquired data.

So far, scholarship has defined and treated data acquisition as a particular step in the data creation process, existing in a vacuum [6, 76, 152, 245, 335]. Very little is known about how this step influences the stages that precede the data labelling, eventually affecting the shape of the final medical dataset. Our studies show that regulatory constraints, the context of data creation and use, and the business models and scalability of the organisations, crucially affect the extent and the type of data that is possible to be collected and processed.

13.5.2 *Conditioning the ground truth design*

Within this context, we identified the design of the ground truth schema as a crucial stage of medical dataset creation. In our studies, the externally imposed constraints shaped the amount and type of data that reached the stage of designing ground truth schema. This has implications for scholarly discussions that focus on developing documentation frameworks that support the responsible and informed use of complex datasets [11, 29, 123, 164, 231]. We showed that the decisions taken during the design of the ground truth schemas were foundational to the succeeding stages of dataset creation. We argue that in this stage, experts do not deal with ideal conditions, but there are inherent limitations which we conceptualised as epistemic differences and limits of labelling. We further argue that the external constraints influence how these inherent limitations manifest in situated collaborative domain settings.

The amount and type of data that reach the ground truth schema design is already shaped by the necessity of organisations to comply with regulatory standards. This has led the experts from ORG I and II to work with data that had limited representativeness from the start, further affected by the predefined purpose of use and geographical, demographic, and linguistic context for its collection and use. These had implications for the negotiations between data scientists, designers, and medical professionals on what "makes sense" to be labelled.

Domain negotiations that we observed, were grounded in epistemic differences that did not take place with symmetrically allocated roles, where the “separation of concerns” of each domain expertise is often negotiated against the tacit medical knowledge but where data scientists have the first say [165, 278, 315]. Having the development of AI models as the purpose of medical dataset creation, data scientists were positioned as the problem owners of the data creation processes. This further distanced the design of the labels from the medical domain experts and was manifested through misapprehensions about medical knowledge and practice. The tensions with the medical professionals often led to negotiations about what was medically important to be annotated versus what would lead to high-quality datasets from a data science perspective. At the same time, both of these standpoints had to correspond to the demands of the health-tech market.

We found that the externally imposed concerns, such as compliance with regulatory standards, the context of creation, and the intended use of the data, along with the commercial and operational pressures, condition the data collection and can affect ground-truth schema design. In fact, many crucial decisions and negotiations relevant to the final shape of the medical datasets take place during the stage of ground truth schema design. All three organisations under study were committed to developing AI systems in a responsible way. As such, the creation of high-quality training data was a crucial step. Yet, no matter how hard they tried to create representative, consistent, well-structured, high-quality data, the resulting datasets were already limited in different ways. We showed how these limits were predefined even before any data labelling occurred. The combination of external constraints that limit and structure data collection with the misapprehension of domain practice resulted in highly paid experts having to imagine and invent additional information to perform the tasks asked of them. A limited understanding of what is required for diagnosing various conditions from medical images could have consequences. Either new datasets would have to be created, which translates into a new data collection process, with all the regulatory constraints attached, or the labelling software would have to be more aligned with the existing professional practices following the guidance of expert annotators. Even where these issues were resolved, medical professionals annotated data based on their particular experience and tacit knowledge. This means that the geographical location of the experts affected what they expected to see in the data, showcasing that expertise does not account for the uneven distribution of diseases in different parts of the world.

13.6 LIMITATIONS AND FUTURE WORK

Our contribution builds on qualitative data from three organisations located in countries of the Global North. Creating medical AI datasets in different countries of the Global South may present different challenges and be influenced by a different set of factors that were not captured in our data. Further research is needed to better understand how medical AI data creation varies across different regions and cultures.

Our study focuses on only two medical areas: radiology and clinical trials. While we engaged with diverse types of medical data, creators of other medical datasets could face challenges unique and dependent on different types of medical specialisations. Future research should aim to explore the factors that influence the design of medical AI datasets across a wider range of medical specialisations to develop a more comprehensive understanding of the factors that influence it.

13.7 CONCLUSIONS

In this paper, we investigated the work of data scientists, medical professionals, and designers that takes place before the labelling of medical data. Building on the qualitative accounts of our ethnographic findings, our main contributions are:

- conceptualising five factors that influence the creation of medical datasets;
- disclosing how these factors condition the design of ground truth schemas;
- suggesting identified relationships amongst these factors;
- staging the design of the ground truth schemas as a highly contested, yet crucial step in the creation of medical datasets that precedes and conditions data annotation.

These overarching factors had a fundamental influence on the final shape of medical datasets created for AI use. First, the externally imposed constraints should be systematically taken into account during the entirety of the medical dataset creation processes, as these factors define the data collection and condition the design of the ground truth schemas. Second, we have exemplified the breadth of decisions taken before the annotation of medical data. Foundational decisions about the final shape of medical datasets take place during the design of a ground truth schema. Future endeavours in data science, law, and policy should consider this stage as crucial to achieving responsible medical AI.

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to all of our participants, especially Dr. Elijah Kwasa, Dr. Edward Mwaniki, Dr. Marian

Morris, Dr. Ruth Wanjohi, Dr. Mary Onyinkwa, Dr. Sayed Shahnur, and Dr. Samuel Gitau for their invaluable contributions and insightful input. Thank you for taking the time to engage with us and for your significant impact on our work.

14

PAPER III: TOWARDS CLINICALLY USEFUL AI

TITLE

Towards Clinically Useful AI: Grounding AI Visions in Radiology Practices in Global South and North

AUTHORS

Hubert D. Zając, Tariq O. Andersen, Elijah Kwasa, Ruth Wanjohi, Mary K. Onyinkwa, Edward K. Mwaniki, Samuel N. Gitau, Shawnim S. Yaseen, Jonathan F. Carlsen, Marco Fraccaro, Michael B. Nielsen, Yunan Chen

VENUE

ACM Transactions on Computer-Human Interaction

SUBMITTED

February 2024

ABSTRACT

Despite advancements in Artificial Intelligence (AI), its clinical adoption remains low, mainly due to its technology-centric development. This challenge calls for designing human-centred AI which takes into account the differences in medical contexts. To inform the design of clinically useful AI in radiology, we conducted a field study of chest X-ray practices with eighteen radiologists and four radiographers from nine medical sites in Denmark and Kenya. During the observations, we asked about their visions of AI support. The findings present nuances of chest X-ray practices and show how they generalise into three stages across countries and medical sites: selecting, interpreting, and reporting. We discuss how HCI can expand the design space of clinical AI by situated future envisioning. Finally, we reflect on how clinical usefulness depends on the configurability and flexibility of AI across three dimensions: type of clinical site, expertise of medical professionals, and situational and patient contexts.

14.1 INTRODUCTION

Artificial Intelligence (AI) has been hoped to significantly transform healthcare by improving patient care, lowering cost, and elevating the work life of healthcare providers [48]. A plethora of studies have substantiated these hopes by demonstrating increased detection rates of pathologies on medical images in controlled environments [210], supporting the detection of cancer on mammograms [126], identifying brain tumours on magnetic resonance images [176], detecting arrhythmia on electrocardiograms [16], or supporting polyp detection during colonoscopy [337]. However, only a small percentage of these systems make it from the lab to the real world [82]. And even if they do, their positive impact on practice and patient outcomes is not guaranteed [26, 208]. Consequently, despite the growing popularity and ongoing technological developments, the successful use of AI in clinical practice remains low. [10, 82, 147, 184, 273, 292, 338, 362].

This has been particularly apparent in radiology, which is one of the first medical fields to confront the potential of increasingly capable AI. There are two confounding factors that make radiology a promising domain to benefit from AI support: (1) the abundance of digital data, the relative use of medical imaging across specialities, and recent progress in vision algorithms performance [312]; and (2) the fact that most countries suffer from severe staff shortages that result in less time available per examination, stress, and overburdening of medical professionals [279]. However, the benefits of using AI in practice have been vague. For example, from 62 models aimed to help with the detection and prognostics of COVID-19 on chest radiographs (X-rays) and thoracic CT, none were identified as clinically useful [280]. Currently, only a handful of systems targeting chest radiography are approved by authorities in the United States and European Union [3] and their clinical utility remains mostly unclear [206].

Current clinical AI development efforts focus primarily on implementing new models and evaluation focused on technical metrics [183, 277], rather than in-situ usage and real-world implementations [313, 340]. This predominantly techno-centric paradigm has contributed to the poor adoption of clinical AI [43, 131, 299, 349, 365]. Prioritising work on algorithms, data, and performance rather than the human and social aspects of AI has been described to result in insufficient and delayed attention to the social considerations of work and, consequently, a mismatch between end-users needs and delivered functionalities [272, 313, 362].

This resembles the beginning of clinical information systems [87], where insufficient understanding of medical work practices resulted in adoption failures [148, 149]. For example, a histology expert system aiming to support pathologists failed due to integration to workflows requiring constant context switching, time-consuming manual data entry, and user resistance [149]. Koppel described how "machine rules that do not correspond to work organisation or usual behaviours" in physician order entry systems facilitated medication errors [193]. Finally, McCauley & Ala called expert systems "a solution (perhaps)

lacking an agreed-upon problem" [222], which highlighted the detachment of motivation for system development from practical challenges [148].

These issues were recognised by the HCI and Computer Supported Cooperative Work (CSCW) communities, who turned towards ethnographic research to understand the organisational context and work practices within entities implementing new information systems [41, 103, 188, 242, 262, 270, 320]. Engaging in formative research within the healthcare domain [23, 51, 78, 111, 157, 174, 256, 257, 274, 296, 367, 368], e.g. through ethnographic inquiries, workplace studies, and user studies, has empowered researchers to design systems addressing the sociotechnical context, human factors, and aligned with end-users needs. For example, Chen identified a functionality gap in Electronic Health Record (EHR) systems and suggested including support for transitional information, which at the time was crucial to the clinical workflow and unaccounted for by the EHR system [78]. Similarly, Zhou et al. discovered that a physician order entry system was not used as intended due to "a change in physical location, sufficient convenience, visibility of the information, and permanency of information". They suggested reframing the requirements for such systems to include both formal and informal work practices [367]. These formative studies transformed the early health information systems from a technological novelty to an integral part of clinical practice [115].

In response to the prevailing technology focus in AI development, the HCI community calls for a turn towards Human-Centred AI (HCAI), emphasising the facilitation of social participation among end-users and the creation of AI that supports self-efficacy, responsibility, and oversight [71, 304, 305]. With AI facing similar challenges to the early clinical information systems, inspired by the rich history of HCI and CSCW engagements, researchers are advocating for studies investigating the sociotechnical context of end-users work in the wild [82, 169, 253, 281, 325, 362], particularly, when designing with such difficult medium as AI [355]. New studies emerge, investigating AI's place in the broader ecosystem of collaborative medical work practices [169, 290, 302, 357], origin of trust [169, 267], explainability [205], bias [18], or collaboration [34]. However, to develop useful HCAI systems, designers and developers have to engage in "more holistic and in-the-wild" methods [10].

This paper contributes to these ongoing efforts of the HCI community by investigating how work practices in medical sites across Denmark and Kenya shape the opportunities for AI support. We conducted a field study through participatory observations in nine medical sites in Denmark and Kenya with four radiographers and eighteen radiologists. Observing first-hand the daily collaborative work and challenges faced revealed significant similarities across Denmark and Kenya in practices and opportunities for AI in medical settings of similar type – defined by the specialisation of provided services and staff.

To explore the connection between the practice of radiologists and opportunities for AI support, we asked our participants how

AI could support their work. Drawing inspiration from prior future-oriented HCI research, we encouraged our participants to be critical towards technology and envision their ideal future [180, 314, 322]. Moreover, we asked our participants about their visions at their workplaces, in the context of their practice. Through grounding in practice, we wanted to ensure that their visions were pragmatic responses to actual challenges rather than hypothetical exercises attempting to retrofit preconceived AI models into their workflows.

This study contributes the following to a growing body of HCI and HCAI research that investigates practice to inform design opportunities for clinical AI-based support systems:

1. Unpacking radiology work practices into three stages across different clinical sites in Denmark and Kenya: selecting, interpreting, and reporting;
2. Showing how HCI may expand the technology-shaped design space of clinical AI and bring it closer to the ideals of HCAI by engaging end-users and ideating visions for AI support at the point of practice;
3. Reflecting on how clinical usefulness depends on the configurability and flexibility of AI across three dimensions: type of clinical site, expertise of medical professionals, and situational and patient context.

14.2 RELATED WORK

14.2.1 *Technology-Centred AI Development*

To some extent, the current state of AI in healthcare resembles the reality of computer science on the eve of the formation of HCI as a research domain in the 1980s. The development of computer systems, in this case, AI-based systems, is claimed, defined, and envisioned through predominantly technology-centric approaches [304, 305, 349, 365]. This state of affairs is partially understandable, as technological progress enabled this new type of system. However, the main drawback of technology-centred AI development is the foundational character of the early decision [6]. Research has shown how the very early decisions taken during data work affect the capabilities of AI models [240, 289, 361]. Going further, the capabilities of AI models determine the design space of potential solutions [237]. Any misunderstandings and misconceptions before and during AI model development would propagate to the final system. Unfortunately, it is often difficult to mitigate them later, as that part of development requires the most effort and resources to conduct [6, 240, 355]. Consequently, the capabilities of an AI model often determine the final system's capabilities. This remains true even for Large Language Models applied to expert tasks, which healthcare arguably falls under [5].

As a result, AI-based systems perform inconsistently when deployed in clinical settings, and their added clinical value is often

dubious [199, 207, 226, 313, 321]. Petitgand et al. investigated the implementation of an AI-based Decision Support System (DSS) in an Emergency Department (ED). They discovered that the system did not integrate with other IT systems and poorly modelled clinical practice. This resulted in support that was technically correct but not useful in practice [261]. Another system evaluated by Hollander et al. intended to support admission to ED decisions due to heart problems [155]. However, the system's output relied on specific cardiac markers available in the hospital's system after a clinician already made the admission decision, rendering the support irrelevant. Similarly, Beede et al. [26] and Wang et al. [334] highlighted issues with AI-based systems developed for work practice that was not reflected in reality, which challenged or outright impeded their use in clinical settings.

Similar issues can be observed in radiology. The almost exclusive focus on detection performance in environments isolated from clinical practice was reflected in the reported challenges hindering the clinical utilisation of AI in radiology [340]. Moreover, AI-based systems in radiology have been found to be often designed for narrowly defined pathology detection cases with limited external validation and disposition to bias [183, 277]. Consequently, when reviewing AI's ability to improve efficiency and health outcomes, van Leeuwen et al. reported that the clinical value of such systems is still unknown [206]. Whereas, Strohm et al. pointed out that one of the main reasons for hitherto AI failure in radiology is the "uncertain added value for clinical practice of AI applications" [313]. This is an opportunity for HCI researchers and practitioners to engage in formative work with radiology practitioners to explore alternatives to the technology-shaped design space of AI-based support systems in radiology and to inform the design and development of AI-based systems focusing on providing clinical value.

14.2.2 *HCI Push for Human-Centred AI*

HCI researchers noticed the challenges of developing useful AI-based systems in healthcare and argued for a shift in AI development to achieve human-centred AI by focusing on human values, responsibility, participation, and oversight [304, 305]. So far, there has been no definitive approach to developing HCAI [71]. Instead, researchers explored different avenues. Trust has been widely recognised as a necessary component of the successful adaptation of AI. Among others, the HCI community informed its clinical and practical origin [18, 62], relationship with accepting and acting on AI output [307], and dependency on organisational accountability [267]. Linked to trust is the "black box" problem - a problem particular to AI, where users cannot inspect and understand the inner workings of a system [72]. Explainable AI (XAI) has been the most promising answer to opening that box, enhancing trust, supporting oversight, but also increasing the perceived usefulness of medical AI [67, 135, 205, 348]. These new

ways of reasoning by and with AI models prompted research into modes and premises of human-AI collaboration [34, 52, 68, 100, 236], investigating reliance, bias, and its potential mitigation techniques [18], problematising AI's authority in such arrangement [179], and its place in the broader ecosystem of collaborative medical work practices [169, 236, 290, 302, 357]. Particularly, the alignment with work practices and sociotechnical context and responding to the actual needs of clinical end-users has been considered crucial when transitioning from technology-centred to human-centred AI [169, 281, 325, 330, 362].

The importance of understanding the clinical context, work practices, and end-users should come as no surprise. HCI researchers showed time and time again that insufficient and delayed attention to the social considerations of work leads to inadequacy of developed systems and inevitable failure in practice [162, 163, 317]. The response to these challenges is one of the most foundational contributions of the HCI and CSCW communities – the push for ethnography-based user and workplace studies to inform system design [23, 41, 51, 78, 85, 103, 111, 157, 162, 174, 188, 218, 242, 256, 257, 262, 270, 274, 294, 296, 320, 367, 368].

The need to understand end-users' clinical context and the sociality of work is just as important now as it was at the beginning of personal computing. HCI researchers describe how formative user studies may serve as an essential foundation for designing and developing clinically useful AI-based systems for pathology detection [136], chronic conditions [314], mental health [326], and collaborative environments of Intensive Care Units [175]. However, conducting meaningful formative studies for AI in a clinical context is afflicted by difficulty in obtaining access, engaging medical professionals, and using AI as a design medium [84, 196, 300, 355]. In the context of still technology-driven AI development, HCI is in need of investigating new ways of involving users in foundational work to shape future systems and make them useful to medical professionals in clinical practice.

14.2.3 *Performance Does Not Equal Clinical Usefulness*

The natural consequence of centring AI development around algorithms, models, and data is evaluation based on technical metrics, such as an area under a receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. The performance-first evaluation approach permeates the general research of clinical AI. Li et al. reviewed the added value of AI when diagnosing thoracic pathologies. The primary metrics reported were sensitivity, specificity, accuracy, AUC, and time spent diagnosing [210]. According to Wang et al., such retrospective cohort studies constituted 98% of all the studies between 2015 and 2019 on AI in radiology [338].

Excelling at narrowly defined technical metrics may not be enough to bring meaningful change to medical practice [37, 182]. Keane et al. argued that none of these metrics ultimately relate to a change in

patient outcomes [182]. For example, breast radiography (mammography) has seen a significant uptake in the use of Computer-Aided Detection (CAD) systems. However, successful use in clinical practice did not improve patient outcomes [208]. This conundrum exemplifies that performance metrics are ultimately relevant during development but do not reflect the benefits of using an AI-based system in clinical practice [303].

This dichotomy has also been noticed by HCI researchers, who have been focusing on human-centred qualities of AI-based systems like trust [18, 62, 169] and usability [67, 195, 237], which support the actual use of a system in practice. However, similarly to the technical metrics, on their own, they do not guarantee a positive impact on clinical practice and patient outcomes. Other researchers investigated what other qualities make an AI-based system clinically useful, which was used as a general term describing positive contribution to clinical practice and patient outcomes [27, 117, 334]. Bossen and Pine found that flexible integration into the clinical workflow, support for sensemaking, awareness of unreliability, practitioners remaining in control, and ability to experiment contributed to medical professionals considering the AI-based tool useful [52]. Similarly, Wang et al. investigated the use of a clinical DSS. They found that the perceived usefulness stemmed from the tool supporting the clinical diagnostic process, facilitating information search, offering training opportunities, and preventing adverse events [334]. These findings are crucial to our understanding of factors influencing the clinical usefulness of AI-based systems. However, to design for clinical usefulness, we need a better understanding of end-users' needs and expectations [326].

14.3 METHODOLOGY

14.3.1 *Study Design*

The goal of this study is two-fold: (1) to inform the design and development of an AI-based chest X-ray support system and (2) to contribute to the understanding of opportunities for AI support in chest X-ray practice across countries like Denmark and Kenya.

To capture the diversity of work performed by radiologists, account for the differences, and identify commonalities, we encompassed various medical sites providing medical imaging services. This field study comprised in-situ participatory observations with radiographers and radiologists in nine medical sites in Denmark (4) and Kenya (5) (Table 15). We grouped the visited sites by the catered population, specialisation level of medical staff, available resources, and size. Two specialised hospitals (D1, K1) provided tertiary and quaternary care, handling the most complex medical procedures in their respective countries. Five general hospitals (D3, D4, K3, K4, K5) offered primary and secondary care and referred patients requiring more specialised care. Lastly, two imaging clinics (D2, K2) provided medical imaging services to patients referred by external physicians

SITE	TYPE	WORK LISTS GROUPED BY	RADIOLOGISTS	COUNTRY
D1	Specialised hospital	Speciality	100+	Denmark
D2	Imaging clinic	Single list	<5	Denmark
D3	General hospital	Modality	<20	Denmark
D4	General hospital	Modality	<5	Denmark
K1	Specialised hospital	Speciality	<20	Kenya
K2	Imaging clinic	Single list	1	Kenya
K3	General hospital	Single list	1	Kenya
K4	General hospital	Single list	<5	Kenya
K5	General hospital	Modality	10	Kenya

Table 15: Visited medical sites.

for chest radiographs, ultrasounds, or CT scans (only K2). K2 also provided teleradiology services to clinics and hospitals in Kenya.

14.3.2 *Participants*

We recruited participants through email and professional contacts of our collaborators; in total - 18 radiologists and 4 radiographers (Table 16). Senior (consultant) radiologists accounted for 15 participants. This means their reports did not have to be approved by another radiologist. Junior radiologists' reports had to be approved by a senior colleague before they could be shared with ordering clinicians. Radiographers, or radiologic technicians, were trained medical professionals who captured medical images, including chest radiographs.

Fifteen participants had previous experiences with various AI-based systems. The evaluated systems covered a range of modalities. Particularly, radiologists from D1 and D4 had DSS for CT thorax at their disposal. These systems offered functionalities like generating a 3D model, correcting motion, segmenting lungs, and detecting findings. Participants working in K5 piloted a DSS detecting selected findings on chest X-rays and were approached by other companies offering AI-based support (see Figure 15). Finally, doctors from K1 and K2 had past experiences with systems detecting tumours in breast mammography, and the single doctor from K3 had past research experience with AI for chest X-rays.

14.3.3 *Data Collection*

Data collection took place from April 2021 to February 2023. We visited D1 in April 2021, D3 and D4 in February and March 2022, K1 - K5 in January and February 2023, and D2 in February 2023. If such an arrangement was possible, we scheduled our visits to observe work on chest radiographs. No personal information was recorded at any time. In cases where particular X-rays were discussed, they

	POSITION	SENIORITY	WORKS WITH CHEST X-RAYS	EXPERIENCE WITH AI IN RADIOLOGY	SITE
P1	Radiographer	Senior	Daily	Yes	D1
P2	Resident Radiologist	Junior	Training dependent	Yes	D1
P3	Resident Radiologist	Junior	Training dependent	Yes	D1
P4	Radiologist	Senior	Daily	Yes	D1
P5	Radiologist	Senior	Daily	No	D1
P6	Resident Radiologist	Junior	Training dependent	Yes	D1
P7	Radiologist	Senior	Daily	Yes	D1
P8	Radiologist	Senior	Daily	Yes	D1
P9	Radiologist	Senior	Few days a week	Yes	D3
P10	Radiologist	Senior	Few days a week	No	D4
P11	Resident Radiologist	Junior	Training dependent	No	D4
P12	Radiologist	Senior	Daily	No	D2
P13	Radiographers	Senior	Daily	Yes	K2
P14	Radiologist	Senior	Daily	Yes	K2
P15	Radiographer	Junior	Training dependent	No	K3
P16	Radiologist	Senior	Daily	Yes	K3
P17	Radiologist	Senior	Daily	No	K4
P18	Radiographer	Senior	Daily	No	K4
P19	Radiologist	Senior	Few days a week	Yes	K5
P20	Radiologist	Senior	Few days a week	Yes	K5
P21	Radiologist	Senior	Few days a week	Yes	K5
P22	Radiologist	Senior	Daily	Yes	K1

Table 16: Participants, ordered by sites.

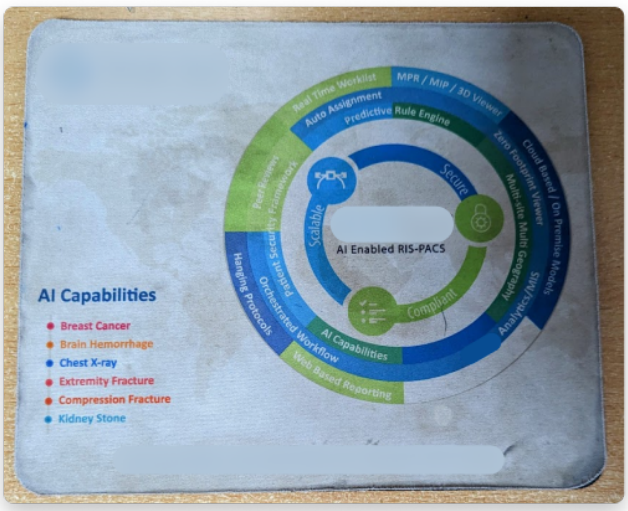


Figure 15: A mouse pad advertising an AI-based system in one of the visited hospitals. The system was not used in the clinic.

were pseudonymised, leaving out personal information, e.g., name and identification number.

We conducted 67 hours of in-situ participatory observations - 35 hours in Denmark and 32 hours in Kenya. The length of observations varied based on the subject and situation in the clinic. This means that our participants decided the times and length of observations. On average, we observed each participant's work for 3 hours and 25 minutes (min. 30 minutes - max. 16 hours). Observations of P1, P9 - P22 were audio recorded, transcribed, and supplied with handwritten notes; during observations of P2 - P8, only handwritten notes were taken.

The main goal of the observations was to learn about the clinic's profile, organisation and division of work at the clinic, patient characteristics and statistics, daily routines, work involved in handling chest radiographs, and past experiences with AI-based systems. Each observation started with informing about the study and observation goals and, if any gaps remained, ended by asking clarifying questions. During the observations, we took place behind the participants' side to observe and ask questions about their work. We halted observations and waited in common areas during meetings or other activities involving patients or unrelated to the focus of the observations. Since we aimed to introduce as little disruption as possible, we engaged in discussion only when our participants indicated readiness by opening up for conversation. Usually, these exchanges occurred during their breaks between examinations or when they encountered something they considered worth sharing regarding the observation goals. Inspired by design anthropological approaches and ethnographic inquiries into possible futures [141, 314], during these breaks, we encouraged radiologists to envision AI support in relation to the work conducted. This approach ensured that their ideas were related to their practice and practical challenges. As a result, the participants generated ideas during the observations, which aligned with their routines.

Participants were not compensated. We collected written consent from all participants. Our study was considered a non-interventional, observational study, thus exempt from a formal ethical review according to the authors' institutions' institutional review boards (IRBs).

14.3.4 *Data Analysis*

Based on the lived experience, information gathered during the interviews, and observation notes, < anonymised pair of authors who conducted the observations> mapped the observed practice of handling chest X-rays by different radiologists across different settings. Next, focusing on high-level outcomes of recorded actions, < anonymised group of authors> iteratively simplified the mapped workflows into a single model of handling chest X-rays comprising three stages: selection, interpretation, and reporting (Figure 16). This workflow model was validated by domain experts - co-authors of

this paper (anonymised). This work was conducted through April and May 2023 in Miro [1] - a collaborative digital whiteboard.

We used thematic analysis [56] to analyse the data collected during the in-situ participatory observations: recordings and handwritten notes. We conducted this analysis in Dovetail[2] - a web application for qualitative data analysis. First, the first author familiarised themselves with the data by manually correcting machine-transcribed audio recordings from the observations and interviews and transferring handwritten notes to a digital format. Second, guided by the gained experience from the empirical work, the first author manually coded any references to work practices, radiographs, and AI (both past experiences and envisioned future use) made by our participants. Third, employing Dovetail's digital canvas, three authors < anonymised> explored the codes for themes. Through weekly meetings through May and June 2023, they iterated the breadth and meaning of created themes in the context of observed practice. As a result, we conceptualised five themes representing challenges encountered in radiologist practice and related visions of AI support.

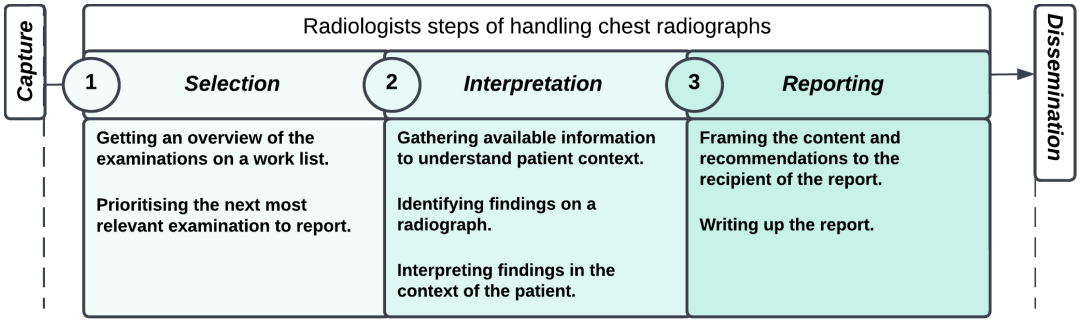


Figure 16: A shared workflow of chest X-ray practice observed in Denmark and Kenya.

14.3.5 *Methodological Limitations*

Several methodological limitations in our study should be acknowledged. First, we did not engage administrative workers who may play a role in determining the distribution of work within the departments. Second, we did not include ordering clinicians in this study. This choice was dictated by the focus on radiologists' practice regardless of the origin of the X-ray orders. Third, we acknowledge that the reporting stage was not explored with as much depth as other stages of the practice. This can be attributed to the fact that the participants' perspective on AI support was largely shaped by their practical experiences and challenges. Fourth, we acknowledge the recent developments and potential of large language models. However, as they were not brought up in the visions for future AI support, we do not have enough basis to address them. Finally, while Denmark and Kenya are two countries from the Global North and the Global South, re-

1 https://miro.com
2 https://dovetail.com

spectively, there is a chance other practices may be observed in other countries, and visions for AI support may differ.

14.4 VISION OF CLINICALLY USEFUL AI SUPPORT ROOTED IN CHEST X-RAY PRACTICE

We expected stark differences between the radiologists' work practices from different countries, primed by the differences in Danish and Kenyan healthcare systems. However, we observed three common steps of handling a single chest radiograph: selection, interpretation, and reporting, present across all the visited medical sites in Denmark and Kenya. Contrary to our expectations, the work practices were remarkably similar in medical sites of similar types across the countries, such as the specialised hospitals (D1, K1), the imaging clinics (D2, K2), and the general hospitals (D3, D4, K3, K4, K5). Notably, the entire process of handling a chest radiograph was remarkably fast. The simplest of radiographs were said to be reported within thirty seconds to five minutes. More complex ones reportedly took between five to fifteen minutes. These estimates included all the work from selecting an examination from a list of radiographs captured in the system to disseminating a report. These stages do not account for all the other activities during a regular workday, like conferences with clinicians, supervising junior radiologists, preparing protocols for CT scans, and performing ultrasounds or CT-guided biopsies. They merely simplify the subset of radiologist work dedicated to handling chest radiographs.

To explore the connection between the practice of our participants and opportunities for clinically useful AI support, we inquired about their visions of how AI could help them at work. Their visions stemmed from practical challenges faced during everyday work and focused on providing actionable support rather than automating some of their tasks (Table 17). This was well-captured by P4, a senior radiologist from a specialised hospital in Denmark (D1), *Make AI that helps me do my work better or faster*. Increasing efficiency and improving patient health outcomes were the basis for every envisioned functionality.

14.4.1 *Selection: Practice of Prioritising*

Radiologists in both countries started their work by logging in to the Picture Archiving and Communication System (PACS), their primary working tool. This is where they selected, viewed, interpreted, and reported examinations. PACS also allowed for accessing historical images and referrals. A referral also called a clinical history or an indication, typically includes a short description of a patient's current state and a clinical question justifying the X-ray order. In cases where PACS was not integrated with a system used by ordering clinicians, the paper-based referrals were either delivered physically to the radiologist (K2), entered to PACS by radiographers when capturing a

SELECTION
Challenge: Backlog of X-rays to report exceeding daily processing capacity
Vision: AI distributing examinations by user's expertise
Challenge: Selecting the next most relevant examination to report without an easy overview
Vision: AI detecting medical emergencies
INTERPRETATION
Challenge: Interpreting visually ambiguous findings
Vision: AI providing decision support on subtle and difficult cases
Challenge: Time-consuming process of obtaining additional clinical information
Vision: AI measuring visual features and comparing changes across historical examinations
REPORTING
Challenge: Conveying the right information in a report
Vision: AI double-checking reports against radiographs for missed or misinterpreted findings

Table 17: Challenges encountered in chest X-ray practice and envisioned AI support.

radiograph (K3), or scanned and uploaded to PACS when capturing a radiograph (K4, K5). PACS was independent of the Electronic Health Record (EHR) system in every visited site, and there was little to no integration between them.

14.4.1.1 *Challenge: Backlog of X-rays to Report Exceeding Daily Processing Capacity*

In PACS, radiologists accessed their work lists for a given day and gained an overview of the examinations to be reported. At that time, the radiographs captured in the evening the previous day or at night were available for reporting. More X-rays were taken during the day. In most of the general hospitals in both countries (D3, D4, K4, K5) and the imaging clinic in Kenya (K2), the number of chest radiographs taken daily often reached or exceeded their reporting capabilities per day, which was voiced by P10, a senior radiologist from a general hospital in Denmark, *I just dream one day. I will open my list and just see ten examinations —– always around or more than 50*. Usually, the number of patients fluctuated depending on the day of the week, season, or proximity of important dates, e.g., beginning or end of school or holidays.

Typically, with several radiologists in a clinic, one radiologist was responsible for a single work list daily. A work list in PACS held examinations of the same type that needed to be reported. The type was defined on the hospital or clinic level and could be based on, e.g., modality or the department of the ordering clinician. This means that radiologists were responsible for reporting examinations assigned to

their specific list, which were usually similar. Sometimes, a radiologist was responsible for more than a single work list, especially in specialised hospitals and during increased patient visits. How radiologists and examinations were assigned to work lists varied depending on the medical site (Table 15).

- *Specialised hospitals (D1, K1):* Work lists were created based on medical specialisations and departments in the hospitals, e.g., infectious diseases, oncology, or heart medicine, and all types of the captured imaging were included as captures as part of their work, e.g., radiographs, CT, or MRI.
- *Big general hospitals (D3, D4, and K5):* The work lists were created per imaging type (modality), i.e., all CTs were assigned to one list, and all radiographs were assigned to another, sometimes with few exceptions, e.g. oncological CTs. In these hospitals, radiologists worked on a weekly schedule, assigned to one of such lists a day.
- *Small general hospitals and medical imaging clinics (D2, K2, K3, K4):* Usually, there was only one list, which could have been filtered for a specific imaging type.

VISION: AI DISTRIBUTING EXAMINATIONS BY USER'S EXPERTISE. Radiologists envisioned an AI-based system that would "filter the normal radiographs" in times of increased workload. It would screen all incoming chest radiographs for findings and assign them to one of two categories: those with abnormal findings and those without. P19, the head of the radiology department in a large Kenyan general hospital (K5) suggested that *the ones that are not flagged as abnormal are given to junior radiologists to clear the dispute.* This means that junior radiologists would ensure that a radiograph is truly normal and correct for errors. P19 continued, *And then the ones that have findings, I would assign to one of the senior radiologists. That would be useful.* This way, senior radiologists would use their time efficiently, focusing only on radiographs with findings. This suggestion sprouted when we asked about the backlog of 500 unreported radiographs. Through such an AI-supported distribution, P19 hoped for more radiographs to be reported within a given period to reduce the backlog to normal.

This suggestion was considered a useful tool to support the smooth work of a radiology department in times of increased workload. However, radiologists were aware of the potential negative effects on the education of junior radiologists should such a system be used daily, which was not their intention. Such a system was envisioned as a temporary means of support in moments of greater stress on the healthcare system *for the purpose of clearing the backlog. 500 [cases] is not too much. Once you get down to a reasonable number, then, of course, they can stop. But it'll be a good filtering tool to clear the queue* [P19]. After clearing this temporary backlog, the distribution of examinations would return to normal to re-engage junior radiologists in education.

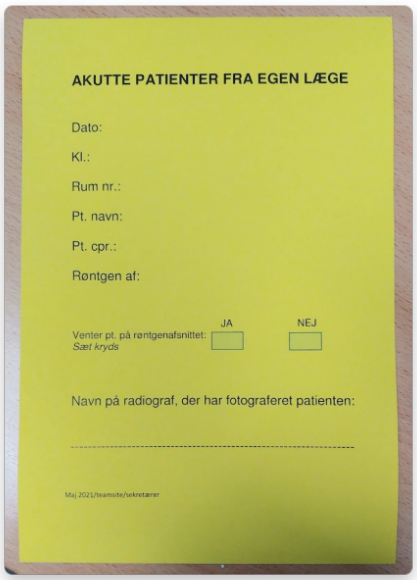


Figure 17: A note from D4 indicating medical emergency. The emergency was indicated by an ordering clinician. A radiographer capturing the radiograph would take note of the emergency and bring this note to a reporting radiologist to prioritise it.

14.4.1.2 *Challenge: Selecting the Next Most Relevant Examination to Report Without an Easy Overview*

When selecting the next examination to report, radiologists in Denmark and Kenya constantly scanned for the next most relevant examination. In the observed sites, medical emergencies constituted urgency and were always prioritised by radiologists. A medical emergency was usually indicated by ordering clinicians on an X-ray order, as they were the ones who knew patients' health conditions (see Figure 17). Importantly, these medical emergencies did not account for accidental findings, i.e., conditions that could constitute an emergency but were not expected by the ordering clinician.

The meaning of urgency varied between the sites and situations. Not all the consequential radiological findings, e.g. cancer, constituted urgency. This was explained by P22, a specialised radiologist from K1, *if you look at the medical emergency, things that need to be actioned, it's not everything. A lung nodule that has been sitting there for a few months, even if I report it on Monday [after a weekend], doesn't make a big difference, but a pneumothorax... makes sense.* This quote emphasises the importance of timely treatment for selected conditions, e.g., pneumothorax, which almost always constitutes a medical emergency, as a delay of even a few days can negatively impact patient outcomes. Whereas prospects with other conditions, despite their seriousness, are less affected by a delay of a few days. Moreover, the conditions prevalent in different hospitals may depend on the type of medical site and the population these sites serve. P17 highlighted the inherent complexity of deciding a priority among other cases that do not constitute a medical emergency. *I can't say this [a single radiological finding]*

is a high priority and ignore the other because it is one scenario, one package. I can't tell which one's more important. In different contexts, different findings may be more or less critical, e.g., the air in the abdomen is an expected finding in a post-operative X-ray but an uncommon one in routine assessment, mandating immediate action in a routine X-ray.

While the utmost priority was given to urgent medical cases, medical professionals, whenever possible, tried to alleviate patient inconvenience. This prioritisation was observed when a patient had to wait for the report, often to bring it to the ordering clinician (K2, K4, K5). Such patients were noted by secretaries or radiographers, who informed radiologists about their situation. This way, the waiting patients' examinations were picked before other non-urgent examinations but after any constituting a medical urgency, as explained by P21: *You see some of these people are not waiting here. They got the exam and went home; others are sitting at the reception. So there's a receptionist there who comes to tell you: "This one is waiting here; please prioritise". So we will do that one and then move on with the rest of the list.*

Medical emergencies were communicated in several ways and, to varying degrees, supported by the existing IT systems. Radiologists typically received communication about the most critical cases through phone calls from the ordering clinicians. This was observed in both countries even when a priority attribute could be indicated on a radiograph order. Such calls were made primarily between doctors within the same hospital, as noted by P17 : *[Even at night] they call us. If there is an urgent [case], they have to call me: "We have an urgent case. Please report it immediately." So I have to wake up and report.* However, maintaining the delicate balance between medical urgency and convenience was difficult. Some phone calls were motivated by the organisation of work at the requesting institution and not a patient emergency. *<After putting down the phone> Urgent thing... or it's urgent because they're closing around five. It's always urgent. So around this time, everyone starts calling* [P14]. In such cases, the phone calls may not always achieve the desired prioritisation but instead introduce breakages to the workflow.

Instead, radiologists preferred the urgency conveyed through PACS as an attribute of a radiograph order. Using this attribute, they could order their work list to select the next most relevant examination to report. In K1, examinations were colour-coded based on the urgency attribute: *The red is an urgent, very urgent examination... So this one you want to report within 1-2 hours, preferably before 1... The next colour will be semi-urgent, so within two hours, preferably before. Then there is this <colour>. Usually, the turnaround time is 24 hours. So the colour codes help us to know which exams we need to report first* [P22]. This way, radiologists could get an overview of their patients, plan their work accordingly, and minimise interruptions caused by phone calls.

Without the prioritisation in PACS, radiologists relied on secretaries and radiographers informing them about patients requiring immediate attention, or established other processes to gain an overview of their work lists and select the most relevant examination next. Not

every PACS in the visited sites supported that functionality (D2, D3, K2, K3, K4), and the system in K5, which did, was rendered unusable due to organisational challenges. Every examination in K5's PACS system was marked as urgent, while almost none of them were. As a result, radiologists ignored the urgency attribute and did not use it for prioritisation. P17, from a small general hospital in Kenya, usually quickly browsed through all the paper referrals, which were scanned in PACS and linked to the examinations, to see which examination to report next.

Lastly, in all of the hospitals, the work and expertise of radiographers were crucial for radiologists' prioritisation of cases. They conveyed the urgency assigned by an ordering clinician in some of the general hospitals (D3, D4, K3, K4). Moreover, whenever assessing the quality of the captured X-rays, radiologists would gain experience in distinguishing certain findings. P9 explained why this was important, *Technicians [radiographers], they were wondering what this was in the breast, but she had both her breasts removed, and I think she had a hematoma, and that's what's remaining... It's good for them to learn because then they can see if there's something urgent, and they will call us*. Based on this experience and patient interactions, radiographers could spot certain conditions on X-rays that necessitated urgency and inform radiologists about them.

VISION: AI DETECTING MEDICAL EMERGENCIES In all the visited sites, radiologists received examinations chronologically and constantly engaged in prioritisation work at the scene to report the most relevant examinations first. However, they had a little overview of the content of the examinations before opening them and relied primarily on clinician-indicated emergencies, which meant that inevitably, some of the urgent findings remained undetected until a radiologist saw the radiograph. Participants from Denmark and Kenya shared similar visions when asked whether they could envision AI supporting them in this prioritisation.

Radiologists envisioned an AI-based system that would screen all the incoming radiographs and detect findings constituting medical emergencies (K1, K2, K3, K5, D2, D4). While the desired findings to screen for varied between contexts and practitioners, pneumothorax - a lung collapse that may lead to respiratory or heart failure - almost always constituted a medical emergency. P14 envisioned AI that would detect it, *you can use it to triage trauma patients, ICU patients, and generally walk-ins... If there's a pneumothorax, it'll let me know. That's a very useful thing.* They would use that knowledge to prioritise relevant cases in their work lists and deliver crucial reports faster. Such screening would improve patient outcomes by faster diagnosis in time-sensitive cases. The main clinical value comes from providing faster diagnoses for patients with undetected medical emergencies.

14.4.2 *Interpretation: Practice of Drawing Conclusions*

After selecting an X-ray, radiologists proceeded to interpret it. The interpretation of chest radiographs involved understanding (1) a patient's situation, (2) visually detecting findings, and (3) interpreting them in the context of a patient's situation, i.e., why the X-ray was ordered, the patient's current state, their location (admitted to a hospital or at home), or their medical history. This results in a high degree of subjectivity. P20, a senior radiologist from a large general hospital in Kenya, warned us that *you can ask each of us [three senior radiologists in the room] and everybody will give a different opinion. That's the biggest problem with chest X-rays* [P20]. We observed two main reasons: the visual complexity of chest X-rays and the variation in the additional medical information required to interpret them.

14.4.2.1 *Challenge: Interpreting Visually Ambiguous Findings.*

Chest radiographs are visually complex, making distinguishing and interpreting single findings difficult. P14 explained how X-rays are merely shadows of the complex structures in the human body cast on a plane. *On the surface, it <an x-ray> is extremely simple, but as you look at it, it becomes extremely complex ... because of all the things that are in the chest and even the different densities within the chest from the lung, all the way to bones, mediastinum, and everything in between.* However, even when spotted, many abnormalities cannot be categorised unequivocally. Shadows of different conditions look the same; some pathologies may hide behind others. For example, during an observation P9 explained, *So I'm looking at this area and wonder if there is anything relevant. But there is something like this on the other side, so it probably is nothing. This is one of the very difficult areas because there are too many bones... and this patient is like this [wrongly positioned], and he's overweight.* P9 explained that they relied on their experience with similar cases to interpret such findings accurately.

One of the main ways radiologists are resolving doubts about their interpretations is through collaborative interpretation with their colleagues (Figure 18). Such collaboration was well-supported by a local PACS in D3. As a result, such consultations became a norm. *I think this is one of the best things about our PAC system. We use it a lot... Instead of calling and saying, "Look at this CPR number [personal identification number]", and she is plotting it in, now this is so easy* [P9]. Later, they encountered an X-ray that they were uncertain about. P9 decided to ask for a second opinion from their colleague: *So I just want to show her this. I don't know what happened, but it doesn't look good at all. Maybe it's nothing.* Then, they concurred at the colleague's desk and worked together to interpret the examination. This effort was deemed worthwhile even when faced with additional labour needed to consult with peers. As the only radiologist in K2, P14 had to rely on collaborating with other radiologists remotely, e.g., *I have a friend who's in the US... maybe two, three months ago, there was one case that... boggled my mind, he specialised in paediatric radiology, but he's also done musculoskeletal ra-*

Figure 18: In K5, the reporting room was occupied by three radiologists at the time of the study. Radiologists often chatted about examinations to resolve doubts. They either turned and looked together at an examination on a single workstation or read out loud an identification number, which other radiologists used to find the examination in question and interpret within their own workstation.

diology so I would consult with him. There's another radiologist who works at (K5) whom I consult with.

Junior radiologists relied even more on collaboration with their senior colleagues. As part of their training, they held regular meetings with their supervisor, which offered an opportunity to ask questions about difficult cases and their interpretations. However, when stuck on a particularly difficult case, junior radiologists would reach out to their senior colleagues outside of the regime of scheduled meetings, as explained by P11, *if I have any questions, I just need to find anyone who is at work [senior doctor qualified to approve x-rays] and I can just confer with them.* These scenarios exemplify the collaborative practices among radiologists from different medical sites and the critical role of such collaboration in chest radiograph interpretation.

VISION: AI PROVIDING DECISION SUPPORT ON SUBTLE AND DIFFICULT CASES Doctors from the majority of the clinical sites (D1, D2, K1, K2, K3, and K5) envisioned two AI-based functionalities that would help them during the interpretation of the radiographs: (1) providing decision support to increase certainty in difficult cases, and (2) detecting subtle finding to reduce the risk of overlooking.

First, our participants wished AI could provide decision support when unsure about their interpretation. While our participants consulted their colleagues and friends when interpreting difficult cases, this support was often burdened with extra work to share the examinations, connect over the internet, or walk to another room to discuss them. P16, the only radiologist at a small general hospital in Kenya

(K3), depicted the following intended use, *I think for me the utility will be more... like a second opinion. I would want it to be like a second opinion to be more confident.* Similarly, P4 explained that in cases when they were not completely sure of a diagnosis, AI providing decision support could help decide on their interpretation, *If I'm 95% sure about the diagnosis, and find something else [an AI prediction] that suggests it, then I'm 100% sure.* While consulting colleagues may be more beneficial, it also takes more resources. Due to the high workload, radiologists often could not read through all the available medical history, let alone consult all the cases among each other. These constraints would not afflict an AI that could opinion a difficult radiograph in question.

Second, they envisioned a system that could help them not to miss subtle findings during the interpretation. *You are often in a hurry. It'll always help if a machine can... show me something I haven't seen. Of course, I can't see everything. Of course, I'll make failures* [P12]. The awareness of the possibility of missing subtle findings was common among our participants. Similarly, P22 from K1 reflected, *As human beings, sometimes you miss something. You miss a small, tiny nodule in the chest x-ray, which after six months... the patient comes back, and it's a big mass which you had missed, which happens. We're humans.* In these visions, radiologists wanted to ensure they had not missed anything by having an AI double-check the radiograph for subtle findings.

These envisioned AI-based systems were contingent on the usefulness of the detected findings. Namely, the need for support was very low when interpreting simple cases with "obvious" findings. This was emphasised by radiologists from Denmark and Kenya alike, e.g., P20 who previously evaluated an AI-based chest x-ray decision support tool mentioned, *when the findings are obvious, it wouldn't even be faster because then... you'll see them quickly. It's with those subtle [findings] that AI could be useful.* The obviousness of the findings was also relative. While most radiologists found decision support detecting pneumothorax extremely useful, a specialised radiologist interpreting post-operative X-rays, where pneumothorax is a relatively frequent finding, held a contrary opinion: *Your tool will most likely detect pneumothorax. It's not that useful. We can easily see a pneumothorax* P7. These comments highlight the lack of clinical usefulness of AI-based systems providing decision support for obvious findings relative to the medical site and professional expertise. With extra work needed to assess and understand these predictions, they offered no value to radiologists or patients.

14.4.2.2 *Challenge: Time-consuming Process of Obtaining Additional Clinical Information*

Chest X-rays were only one of many diagnostic tools medical professionals in Denmark and Kenya employed to understand patient conditions. Insights from chest X-rays were only as good as radiologists' understanding of patients' context. Radiologists in both countries familiarised themselves with a radiograph referral to understand the clinical questions associated with a particular X-ray. Then, they gath-

ered other available clinical information, e.g., historical images, to create a coherent story about a patient's condition. Only then, in the light of their understanding of a particular patient's situation, they interpreted the radiological findings seen on an X-ray. P9 pointed out the broader focus of radiologists' work: *It's all the big picture, and I can't do a proper description of the picture if I don't have all the information*.

A referral was typically available in digital (as a scanned document or a digital entry) or paper form. Radiologists used referrals to focus their efforts and guide their interpretation of visually similar findings. P17 stressed out that *if they [ordering clinician] didn't write it, I don't know what they look for, why this patient did the chest X-ray. Are we looking for an infection or pneumonia? But the patient had a trauma, so we are looking for a fracture. Maybe I would have missed it if they hadn't told me.* For this reason, all of our participants considered a referral necessary information and often would not report an X-ray without it.

Historical radiographs were another important source of information. Comparing patients' current X-rays against previous ones was useful to exclude potentially life-threatening diagnoses and discern between visually similar findings. P12 , from an imaging clinic in Denmark (D2), captured this by referring to old X-rays as gold: *they are gold because when you can see he [a patient] was seen five years ago, you find the old picture and it looks exactly the same. Then you say, "Oh, it's not cancer".* In this scenario, P12 found the same finding of the same size on a historical X-ray. Hence, they could rule out cancer. Similarly, P21 from a large general hospital in Kenya (K5) used historical X-rays to guide their interpretation of ambiguous findings: *The images can look alike. But interpretation will differ depending on the history* [P21]. While providing invaluable insights, fetching historical images and comparing the development of findings across the radiographs was repetitive, time-consuming, and poorly supported by PAC systems. *It takes a lot of time to open one [radiograph], look at it and see what was happening, then open the next one and see if it had resolved.* [P22]. The need for extra time to use historical radiographs was a concern during times of heavy workload.

Medical records provided a wealth of information, e.g., test results or treatment progress, yet radiologists had to go out of their way to access them. Since radiologists worked in PACS and could access the referral and historical images directly, switching systems to search for patient information in EHR was usually time-consuming. Because of this, it was used as a source of last resort for most complex cases. P10 explained: *Not all of us will do this because we do not have enough time. I can't check every patient in the SP [the local EHR system].* However, information obtained in such a way was considered very helpful, as expressed by P22: *I prefer to check the records cause you're able to get even more [information] than the clinician may give you.* EHR was available in Danish hospitals, in the specialised hospital (K1), and the bigger general hospital (K5) in Kenya. Paper versions of the records were not used.

Finally, despite the overwhelming subjectivity of the interpretations, radiologists pointed out that they also use some objective met-

rics. Frequently, radiologists manually measured ratios or sizes of visual features on a radiograph to discern certain conditions. For example, they measured the width of the cardiac silhouette and the thoracic cavity. A ratio greater than 0.5 suggested cardiomegaly (enlarged heart). Similarly, P14 explained, *I am trying to discern whether this area is normal vascularity or whether it's an infection. An objective way to do that is to compare the intercostal spaces.* While measuring such features was common and not difficult for radiologists, it takes their valuable time away from using their expertise.

VISION: AI MEASURING VISUAL FEATURES AND COMPARING CHANGES ACROSS HISTORICAL EXAMINATIONS To reduce the probability of errors and streamline their interpretative work, radiologists envisioned AI-based systems that could (1) pre-process radiographs by measuring visual features and (2) compare detected findings against historical radiographs.

First, AI could assist with manual measurements of visual features. P22 envisioned an AI-based system that could support their interpretation by conducting such measurements: *AI could just measure for me the heart size and tell me if it's within normal. Because it's very manual, time-wasting... But you know, you can easily train a machine to do that.* This way, radiologists would only have to inspect the results, which would realistically reduce the time needed to interpret a radiograph.

Second, AI could automatically compare changes between historical radiographs. Similarly to the manual measurements, fetching historical images and assessing their differences was a common and time-consuming task. Participants from K1, K5, D1, D3, and D4 envisioned an AI-based system that automatically fetched historical radiographs, detected relevant findings using image recognition, and compared their development over time. Providing such support would expedite the manual search for findings across historical images and help radiologists assess whether changes currently observed are new developments or have been already present, resulting in improved clinical outcomes. Rather than spending time on manual fetching and comparing historical radiographs, P22 , working in a specialised hospital in Kenya, envisioned: *It would help to have an AI that just compares... say the last five scans... highlights those changes for you so that you can just interpret what has happened. So that would be nice.* In this vision, the AI-based system would use image recognition to detect changes between historical chest radiographs and bring them to the radiologist's attention for interpretation. P9 , from a general hospital in Denmark, envisioned a similar support tool to find the same finding across historical images and inform radiologists about its progression. *There are some small nodules in the right lung, and I have to compare... I think artificial intelligence could be very helpful here to spot the nodules and then compare them.* The envisioned functionality would allow radiologists to focus on interpreting the changes rather than fetching images and localising the changes.

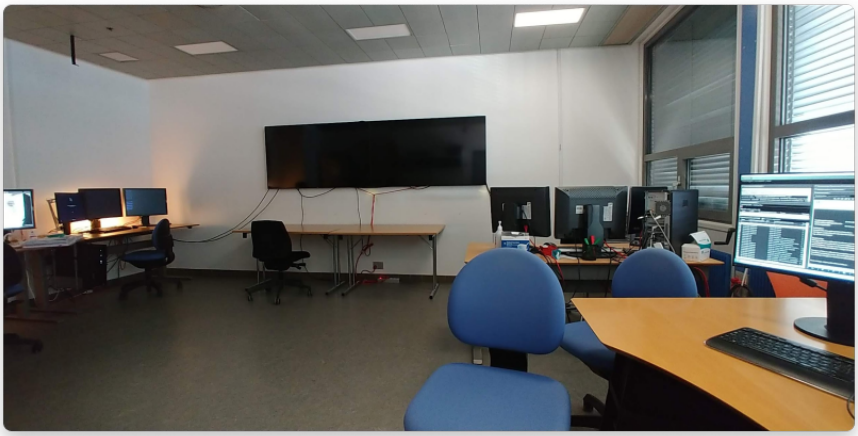


Figure 19: In D1, conferences with clinicians were a part of the dissemination practices aimed at increasing the knowledge transfer between radiologists and ordering clinicians. In this room, radiologists responsible for one work list presented and explained the most important, difficult, or interesting cases to clinicians from the departments that ordered them. This was also when they could ask questions about the reports of other patients.

14.4.3 *Reporting: Practice of Contribution*

The final stage of handling a chest radiograph encompassed writing a report. A report may seem to be solely a plain summary of an X-ray interpretation. This part of radiologists' work was also mediated by a patient's context and the expertise of the report recipient. In the visited sites, radiologists usually communicated the three following elements in their reports: their impressions of what they saw on the radiograph, their interpretation of these findings in relation to the patient's condition, and recommendations on the next steps for the ordering clinician (Figure 19).

14.4.3.1 *Challenge: Conveying the Right Information in a Report*

Radiologists framed their reports to provide maximum benefit to the patient based on the ordering clinic and the report recipient. The level of detail in their descriptions was finer if they knew the ordering clinician could use this information. They also mentioned all the visual findings that may concern ordering clinicians to avoid confusion, even if they were not clinically relevant. *There are some [findings] that I can ignore mentioning. But the others not... they [ordering clinicians] might think that you missed it because they're not aware what it is, they might worry that maybe it is something when it's not* [P16]. Similarly, the type of guidance radiologists provided varied between the recipients. P14 from a medical imaging and teleradiology clinic in Kenya mentioned, *My reports vary depending on who has sent it [an X-ray order]. So if it's a medical centre in Nairobi, I know they're going to interact with a healthcare practitioner who's fairly experienced... but the further out you go out, the less you have...* P14 explained that clinics in the remote parts of Kenya lack medical doctors. As a result, nurses and clinical officers take care

of patients. However, they may lack the necessary medical training to navigate the complexities of diagnosing certain conditions. P14 continued, *I need to use my expertise to sort of guide them. Cause they'll put a lot of weight on what I say... You want to give them some ideas: "This is what I think is going on. Please check for 1, 2, 3." And then they may not have thought of it and said: "Oh well, the child doesn't have a fever, and the child hasn't been coughing, so it's unlikely to be pneumonia." But if I write pneumonia, they're going to put the child on antibiotics whether or not the child has symptoms. So they say, "The doctor from Nairobi says pneumonia, so it's pneumonia." It's tricky...* Overall, these quotes underscore the adaptability and responsibility of radiologists to provide tailored guidance in their reports based on the specific needs and capabilities of the recipients they report to.

VISION: AI DOUBLE-CHECKING REPORTS AGAINST RADIOGRAPHS FOR MISSED OR MISINTERPRETED FINDINGS As a response to the problem originating at the interpretation stage - the complexity of reporting chest X-rays and the high temporal cost of accessing support - and the reporting stage - the fallibility of transcription software used, radiologists envisioned a system that would double-check whether all the relevant findings detected on the X-ray were mentioned in the report. This system would use two types of AI: Natural Language Processing (NLP) to read through a radiologist's report and image detection to detect relevant findings on an X-ray. Subsequently, the system would assess whether all the findings detected on an X-ray were also present in the accompanying report. In case of divergence or missing a finding, it would notify radiologists about potential errors.

Radiologists acknowledged their fallibility, which could be minimised with such AI use as envisioned by P17 , *Even for us as consultants... everyone is making mistakes, right? So I can report everything, and just with one click, I can... see if I was wrong, if I missed something, if I have to think more about this, if I have to ignore this. It'll help everyone.* The report analysis is the main difference between this vision and the decision support described before. In this envisioned future, the AI's assessment of the radiograph is not available to radiologists before finishing the report. Thus, radiologists benefit from seamless quality assurance and interact with the AI only when an issue with the report is discovered, avoiding another layer of work for every radiograph.

14.5 DISCUSSION

This study offers a unique perspective on radiology work practices and AI opportunities in two diverse countries. HCI researchers suggest that such joint focus on commonalities is necessary to balance the northern narrative in technology design, minimise biases, and "enable translation across geographies" of clinical AI-based systems [168, 201]. If not addressed, sociotechnical and political differences both within and between Global North and Global South countries im-

pact the design and successful use of clinical IT systems [26, 168, 248, 252, 342]. We observed such differences in, among others, the type of healthcare systems, IT infrastructure, available IT systems, and the quality of medical data. While their effect on concrete AI implementation should be assessed, the radiologists' practices in Kenya and Denmark were surprisingly similar. The reasons may be due to comparable medical education where the biomedical model of medicine, scientific constructs and medical ethics are largely the same.

We contribute to this line of work by showing how similarities between chest X-ray practices in Kenya and Denmark can be generalised across three main stages of work that are more collaborative than expected and possess unique human contributions. We also detailed the ways in which local and contextual aspects of radiology work are important for visions of AI support. These findings have implications for AI in radiology but also for HCI and the design of human-centred AI. Finally, we discuss these lessons and share reflections on what this means for designing clinically useful AI.

14.5.1 *Practice-grounded Envisioning of AI Futures: Searching for New Design Opportunities*

Previous HCI studies revealed a disconnect between the functionality provided by AI-based systems and the functionality needed in clinical settings [169, 175, 267]. This suggests gaps in understanding the problems and needs of clinical end-users before considering AI-based solutions. To bridge this gap, researchers in the HCI community are increasingly voicing their concerns and advocating for an AI-paradigm change going from technology-centric to human-centric [10, 42, 281, 304, 305, 344, 359]. Results presented in this study adopt a human practice-centric perspective on AI support. We have demonstrated the potential to uncover new ways for AI support that align more closely with the needs and wishes of clinical end-users.

HCI researchers should envision AI functionality through engagement with real-world practices rather than merely following the trajectory of AI development. Engaging communities of practice before the development of AI systems is critical [175, 236, 314, 326], as once completed, AI model capabilities often dictate the final system's capabilities and, due to the high cost of the data and model work, they are unlikely to change [6, 362]. We advocate for a paradigm shift where clinical support opportunities arise from the practice, and not from retro-fitting already conceived AI models born out of access to data or external pressures to adopt AI. Practice-grounded envisioning of AI futures can reveal new ways these systems can support practitioners. Thanks to such engagements, in this study, we revealed opportunities for support in the complex, nuanced radiological workflows and collaborative dynamics inherent in clinical practices. Therefore, as AI continues to increase its presence in healthcare, it is essential to shift the driving force behind AI development, from a technological opportunity to a desirable vision of the future rooted in the

local sociotechnical context (local practices, hospital settings, expertise of clinical end-users).

14.5.2 *New Design Opportunities for Clinically Useful AI in Radiology*

The current understanding of handling chest X-rays has centred around an individual radiologist's interpretation work. This is reflected by hitherto AI-based support systems that have so far been aiming at supporting the interpretation and detection of findings on X-rays, also referred to as second-opinion [89, 183, 189, 277, 340]. However, the clinical use of such systems in radiology remains low [147, 292, 338]. Research points out that one of the main reasons for the failure of AI in radiology is the "uncertain added value for clinical practice of AI applications" [313].

Our study reveals that the low uptake of AI in radiology may be due to the application of AI in a confined design space which can lead to missed opportunities for providing clinical value with AI. The extent of visions for AI presented in this study brings attention to the importance of broadening the design space to include all three stages of chest X-ray practice, transcending the narrow potential of AI-based second opinion. This conceptual contribution is not merely speculative; it is grounded in the insights of radiologists who advocate for a pragmatic shift towards support mechanisms that enhance their professional expertise, as opposed to replacing them with models simplifying their practice [200, 209, 323, 351]. Our contribution lies in highlighting new directions for designing clinically useful AI in radiology and recommending improvements in how AI can support radiology work.

AI support has to account for the subtle collaborative practices in chest X-ray practice we showed the collaboration between radiologists and other actors involved in the treatment process of patients. Fridell et al. previously argued that radiologists are active partners to clinicians in the diagnostic process [120]. We expand this understanding by describing the subtle collaboration often mediated through various artefacts such as referrals, urgency attributes, and other data associated with X-rays. These findings contrast the dominant understanding of radiological practice, which, based on current AI-based systems offerings, focuses on the visual detection of findings on radiological examinations almost out of clinical context. We argue that to support radiologists in their practice, designers and developers of AI-based systems have to be aware of the, in fact, collaborative work practices. For example, the collaboration on prioritisation of relevant examinations. To truly contribute to radiologist work and have a positive effect on patient care, future AI-based systems must account for the already existing practices rather than ignore them, as these practices enable radiologists to work and support the best clinical outcomes for patients.

Meaningful AI support should not simplify the practice to one-dimensional finding detection. Our study foregrounded the human

aspect and the range of tasks in radiologists' practice. Previously, Pespane et al. highlighted, among others, educating students, communicating diagnoses, and performing interventional procedures [260]. We add to this tasks related to handling medical images, e.g., drawing from various data sources to understand patient situations, anticipating ordering clinicians' questions, and framing reports to contribute to clinicians' decision-making processes. These tasks are radiologists' contributions to the diagnostic process beyond the sole identification of findings on examinations. They require empathy and complex reasoning. This is a second account where the radiological practice is miscomprehended by hitherto AI-based systems. Otherwise, to use AI output in practice, radiologists would have to spend additional resources, which the support often aimed to free in the first place. To be able to contribute to radiological practice, designers and developers alike should be aware of their complexity and take them into account when designing support functionalities.

To provide useful second opinions, AI needs to limit additional work needed to use it by adjusting to local conditions. Our findings do not negate the premise of supporting image interpretation. We argue that AI, as a second opinion, should be configurable to the local conditions, including the type of clinic, radiologist expertise, current workload, and patient characteristics, to reduce the additional work required to benefit from the predictions. Our participants stressed the difficulty and subjectivity of interpreting chest X-rays, in line with previous research - the disagreement rates among radiologists on chest radiograph interpretations reach as high as 30% [113]. However, providing a second opinion on every examination based on an arbitrary set of pathologies, as experienced by some of our participants, resulted in a significant overhead. We know that the mental and temporal cost of discerning false positive AI predictions may result in the failure of AI in clinical practice [18, 25, 221, 325]. This is critical in radiology where staff shortages and the increasing number of examinations result in reduced time available per examination, stress, and overburdening [146, 279]. We argue that to provide a clinically useful second opinion, radiologist must be able to configure AI to reduce the additional work and spend their time on predictions relevant to their practice.

AI support for radiology can transcend second-opinion. In this study, we explored the space of AI support thought clinically useful by radiologists. By contextualising chest X-ray practice within the broader diagnostic process, encompassing all its interdependencies, we uncovered several new opportunities for AI assistance. Particularly, we point to the selection and reporting stages as promising domains for AI designers and developers. Within these stages, radiologists envisioned AI to provide a much-needed overview of urgent cases, optimise case distribution based on the current clinic workload, or double-check reports. Interestingly, none of the AI-based systems encountered by participating radiologists explored these practical tasks. Promising work on other modalities has been made to support the work in the selection stage. For instance, Arbabshirani et

al. [13] evaluated a system that screened and prioritised intracranial haemorrhage on CT scans, significantly reducing its time to diagnosis. However, such work for other modalities has not been described yet. This underscores the need for diverse support approaches, as the role of radiologists extends beyond pathology detection in medical images [260].

14.5.3 *Clinically Useful AI Requires Configurable And Flexible AI*

AI accuracy has historically been the main measure for evaluating the usefulness of AI systems in healthcare [207, 210]. Achieving high AI accuracy has been the ethos of what entails a 'good' AI model, resulting in the pursuit of getting technology 'right' before anything else. However, technical excellence does not guarantee a positive impact on patient outcomes or the work of medical professionals [313, 340]. Researchers from both HCI and Health point out that technical excellence rarely translates directly to clinical usefulness [37, 68, 169].

14.5.3.1 *Clinic Type and End-User Expertise Condition AI Support*

Different types of clinical sites demand different types of AI support. This study revealed unexpected parallels in radiology work between Denmark and Kenya. However, while radiology work exhibited general similarities across the two countries, nuanced yet important differences appeared when comparing across the types of medical sites. For example, in specialised hospitals in both Denmark and Kenya, X-ray work lists were created based on the hospital departments and according to medical specialisation, which stood in stark contrast to small general hospitals, where there was usually only one work list. These organisational differences affected the opportunities with AI for each medical site. For instance, the AI-based distribution of examinations by expertise was only envisioned in medical facilities with diversity among the employed radiologists. Similarly, differences in available resources in the radiology department were reflected in the envisioned AI support. In smaller hospitals with a shortage of radiology staff, obtaining a second opinion was harder, which deemed visions for AI-based interpretation support more desirable. This suggests that the characteristics of the organisational context are important to consider when designing AI support in radiology. It highlights the fact that different types of medical sites require different types of AI support.

Junior and senior radiologists have different needs for AI support. Besides organisational differences, we also found that differences in radiologists' expertise may condition specific user requirements for AI. For example, on visually complex X-rays where shadows of different conditions looked the same, radiologists relied on their experience with similar cases to accurately interpret the findings. Similarly, junior radiologists who were undergoing training relied on collaboration with senior radiologists, especially when examinations were difficult to interpret. This suggests that end-users level of expertise

is a deciding factor with regard to which type of AI is considered useful. This, we argue, has several implications for the design and development of AI in radiology.

Clinician-facing AI needs to be configurable to the clinic and end-user. It is common knowledge, both in academia and the software industry, that favourable outcomes can be attributed to the configurability of clinical information systems [151, 212]. While designing for system configurability and tailorability is somewhat old news for HCI, it has yet to become part of the discourse on designing clinically useful AI. Emerging HCI studies investigating clinical AI indicate this direction. Researchers emphasise the significance of poor workflow integration, highlighting its equal importance alongside the well-known challenge of establishing trust by means of transparency and explainability [52, 169, 334]. Varma et al. [330] ascribe the limited impact of AI in medical practice to the poor connection between AI capabilities and clinical workflows. They argue for the importance of diverging from the existing "one-size-fits-all" paradigm within HCAI discourses. Similarly, Wang et al. [334] report on tensions with the design of AI-based clinical decision-support systems in a rural clinical context and stress the reason being the misalignment with local context and workflow. Solutions to workflow integration problems with clinical AI have revolved around human-AI interaction, with extensive work on guidelines present (see e.g. [8]) but also suggestions of creating multi-user systems and designing for time-constraint medical environments have been proposed [169]. When reflecting on the results of this study, we suggest that designers need to look beyond the interaction components of the interface. Specifically, there is a need to carefully consider how to make clinical AI systems configurable so that the set of AI features corresponds to the local needs of the medical site. This includes designing functionalities that can be adjusted according to e.g., the division of work and the number of medical staff. This configurability should ideally extend to the expertise of the end-users. This will, if done well, ensure a better fit between human competencies and the capabilities of AI.

14.5.3.2 *Situational Circumstances and Patient Context Influence Requirements for AI*

Finally, radiologists rooted their visions for AI support in concrete situations and challenges from their local practice. For example, in time-limited situations where radiologists were "in a hurry," the demand increased for AI that double-checks the radiographs for subtle findings or screens a list of examinations for medical emergencies that should be treated before the end of a radiologist's shift. Similarly, the type of welcomed support depended on patient context, e.g., notifying radiologists about the detection of air in the abdomen may be a crucial safety feature when dealing with outpatients and a superfluous hindrance when applied to patients from a surgical ward. This means that radiologists' needs for AI are not constant; rather, they

may change over time according to the situational circumstances and patient context.

Designing flexible AI is critical for ensuring clinical relevance. Bossen and Pine [52] have similarly found that AI being "flexible" is an important factor for the successful use of AI in healthcare contexts. In their study of a Natural Language Processing-based tool in the wild, they found that AI's utility was derived from its ability to act as an imperfect assistant that could be appropriated in use to fit particularly well with individual needs. While the timing of AI is important, i.e. it should be available when needed, HCI researchers suggest interactive machine learning as a way to make AI more useful for "in-the-moment diagnostic needs" [67]. Cai et al. [67] provide important design recommendations for ways to improve AI's clinical relevance by enabling pathologists to interact with the AI tool in ways that fit with the particular cases. In both studies, it is clearly demonstrated how erroneous and imperfect AI algorithms can be made more clinically useful if designed for flexible use accommodating the needs of the situation. In light of these findings, we advise researchers and designers of clinician-facing AI to identify situational requirements and design for the appropriation of AI during use.

14.6 CONCLUSIONS

In this article, we present findings from a field study in nine medical sites in Denmark and Kenya. We unpacked work practices shared by participating radiologists from both Denmark and Kenya and conceptualised them in three stages: selection, interpretation, and reporting. Radiology work was found to be more collaborative than anticipated by being part of the overall diagnostic work. Moreover, the unique human contributions of radiologists surfaced as important yet often omitted when designing AI-based systems for chest X-rays.

Findings from this study suggested a misalignment between the dominant technology-centred development of AI and the contextual needs of radiologists. By investigating work practices and engaging radiologists in envisioning AI futures at the point of practice, we expanded the design space of AI in radiology. This includes visions for AI-based distribution of examinations, measurements of visual features, assessments of historical changes, and double-checking reports. These visions transcend the traditional second-opinion systems and suggest that more opportunities for AI support that target other stages than interpretation should be explored.

Finally, we discussed how the clinical usefulness of AI is dependent on its configurability and flexibility with regard to the type of clinical site, expertise of medical professionals, and situational and patient context. These reflections have implications for the design of clinician-facing AI and suggest new directions for future research on human-centred AI in health.

15

PAPER IV: "IT DEPENDS..."

TITLE

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

AUTHORS

Hubert D. Zając, Jorge M. N. Ribeiro, Silvia Ingala, Simona Gentile, Ruth Wanjohi, Samuel N. Gitau, Jonathan F. Carlsen, Michael B. Nielsen, Tariq O. Andersen,

VENUE

ACM SIGCHI Conference on Designing Interactive Systems 2024

SUBMITTED

February 2024

ABSTRACT

Artificial Intelligence (AI) models repeatedly match or outright outperform radiologists in narrowly defined detection tasks. However, real-world implementations of radiological AI-based systems are found to provide little to no clinical value. In this paper, we explore how to design AI for clinical usefulness in different clinical contexts. We conducted 19 design interventions with 13 radiologists from 7 clinical sites in Denmark and Kenya. The interventions centred around the iteratively improved prototype of an AI-based system. We conceptualised four technical dimensions of radiological AI that, to achieve clinical usefulness, must be configured in relation 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.

15.1 INTRODUCTION

Artificial Intelligence (AI) models repeatedly match or outright outperform radiologists in narrowly defined detection tasks [14, 244, 268, 283]. There are multiple studies claiming that AI-based systems enhance radiologists' work, either by increasing accuracy or reducing time spent on each examination [210]. 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 [362]. For example, Roberts et al. [280] 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 [3] revealed that their clinical impact when integrated into practice remains mostly unclear [206, 340]. A similar study by Lehman et al. [208] 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" [313]. 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 [207, 334, 362]. 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 [37, 184, 216]. Third, end-user acceptance, supported by qualities like trust and usability, emerges as pivotal for successful use in clinical practice [62, 67, 169]. 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 in-

terventions (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 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.

15.2 RELATED WORK

15.2.1 *Clinical Usefulness of AI Systems in Healthcare*

The hitherto evidence of AI's positive influence on clinical practice is limited [207, 226, 321]. Research on the real-world effect of AI in healthcare tends to be discrete, focusing on confined goals [362]. However, to provide clinical value AI-based systems have to dovetail contributions from Human-Computer Interaction, AI, and Health into a cohesive vision [117, 334, 362].

First, clinical usefulness necessitates robust performance [362]. 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 [26]. Poor performance also leads to increased workload [261, 285, 334], additional time spent on discerning false positive predictions [301, 302], or breakages to work routines [122]. Van Leeuwen et al. [207] 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 [184]. However, randomised clinical trial (RCT) - a focused, systematic, rigorous, and insulated method commonly used to evaluate the validity of clin-

ical interventions independent of external confounders - is often following the traditional sequential paradigm of work characteristic for drug development [43]. In this tradition, the intervention is evaluated only when deemed complete [70]. 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 [216]. While technical performance is the backbone of useful AI, clinical efficacy is not its immediate consequence [37, 303]. For example, Lehman et al. [208] 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" [88, 184]. 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 [18, 62] and dependencies [267], as well as issued recommendations for design [169]. Explainable AI (XAI) has been the most promising answer to enhance trust, support oversight, but also increase the perceived usefulness of clinical AI [67, 135, 205, 348]. AI as a new source of information and agency prompted the exploration of new ways of reasoning and human-AI collaboration [34, 52, 68, 100]. Researchers also investigated AI's position in a clinical decision-making process [179] and the rationale behind integration opportunities into clinical practice [169, 290, 302, 357]. 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 [82, 169, 253, 281, 325, 362]. 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 [43, 362]. We need to investigate how AI-based systems can be configured to support these three goals and ultimately result in clinically useful AI.

15.2.2 *System Configurability*

Configurability has been long considered crucial to the appropriation of IT systems [101, 102, 198]. There are two types of configurability that should be explored in the context of this study: before-use and in-use [151].

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 [151]. Various methods and approaches have emerged to facilitate meaningful engagement with end-users, such as participatory design techniques [198]. Acquiring an understanding of work practices and work environment, but also technology aspects of a future system and changes it may introduce, is crucial for developing systems that effectively respond to user needs [186]. 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 [311], 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 [319].

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 [347]. 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" [19]. Dourish [102] 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.

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 [362]. HCI researchers investigating the design of AI-based systems learned that it is not possible to envision all aspects of clinical AI-based systems before they are deployed. As a result, the final capabilities of such systems only take shape after they have been deployed. [128, 354]. On the opposite end of AI innovation, i.e., prior to data labelling, Zając & Avlona [362] 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. [107] 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 [109] who emphasise that

#	TYPE	RADIOLOGISTS	COUNTRY
K1	Small General Hospital	<5	Kenya
K2	Small General Hospital	1	Kenya
K3	Imaging / Teleradiology Clinic	1	Kenya
K4	Specialised Hospital	<20	Kenya
K5	Big General Hospital	10	Kenya
D1	Specialised Hospital	100+	Denmark
D2	Imaging Clinic	<5	Denmark

Table 18: Clinical sites included in the study.

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. The emergence and propagation of dependencies 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.

15.3 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. The project is a multidisciplinary collaboration between <anonymised university>, <anonymised hospital>, and <anonymised private partner>. 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 18), 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.

#	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
I08	P08	Junior	D1	70min	II
I09	P09	Junior	D1	50min	II
I10	P10	Senior	D1	30min	II
I11	P11	Senior	D1 / D2	30min	II
I12	P08	Junior	D1	60min	II
I13	P10	Senior	D1	30min	III
I14	P12	Junior	D1	80min	III
I15	P11	Senior	D1 / D2	45min	III
I16	P10	Senior	D1	30min	III
I17	P13	Junior	D1	40min	III
I18	P05	Senior	K5	95min	III
I19	P04	Senior	K4	70min	III

Table 19: Participants that took part in the study.

The participants were recruited through email and the professional networks of our project members. Nine senior (consultant) radiologists and four junior (in-training) radiologists joined the study (Table 19). Junior radiologists' reports must be 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.

15.3.1 *Research-Trough-Design: Design Interventions with working prototypes*

To explore the clinical usefulness of AI in different radiology contexts, we undertook a research-through-design approach [369]. 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. 20). 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

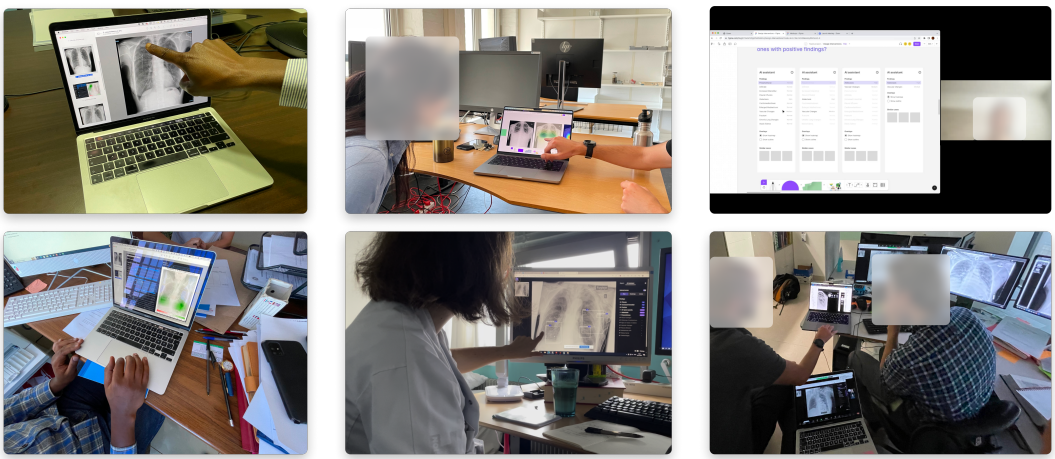


Figure 20: Online and collocated design sessions and design interventions with user interface mock-ups and working prototypes (version I, II, II).

design intervention, as defined by Halse and Boffi [140], is a method that integrates design and ethnography and "enables new forms of experience, dialogue, and awareness about the problem to emerge" (see also [45, 47]). 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" [9].

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 seven design sessions and eleven design interventions (Table 19) 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. We also developed 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 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 (<anonymised>). 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.

15.3.1.1 *Prototypes*

As part of the greater project, a deep learning-based model was developed to detect selected radiological findings <references left out for review>. The first prototype was merely its proof of concept, i.e., it was trained on a small subset of project data, and the final architecture of the model was not finalised. The project was granted access to fully anonymised chest X-rays, which were used in the prototypes. Importantly, it was not a prototype designed for the purpose of collecting 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 [362], we considered it an opportunity to engage in more concrete discussions on the merit of clinical usefulness with medical professionals.

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.

15.3.2 *Analysis Positionality*

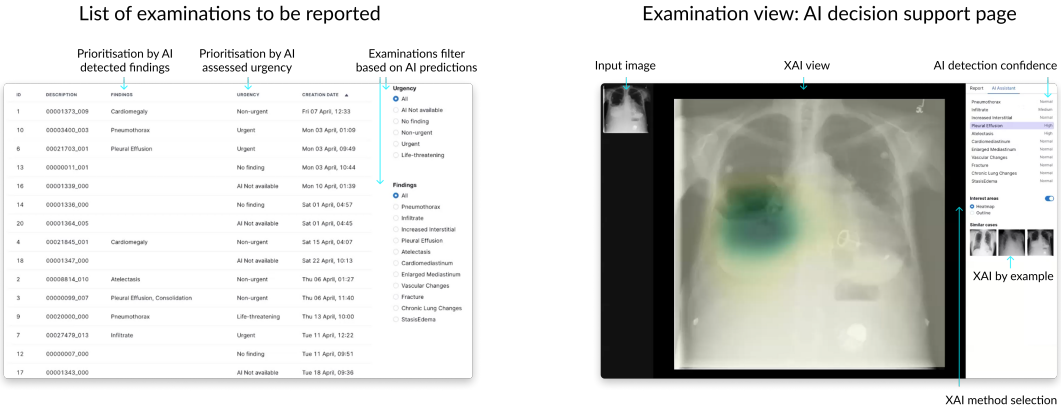
The data analysis was conducted by <anonymised> with backgrounds in Healthcare Informatics, HCI, and AI (5+ & 15+ years of experience). Moreover, before the analysis of the data from design interventions, the two co-authors concluded extensive ethnographic investigations into the work practices of radiologists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

15.3.3 *Data Analysis*

We used reflective thematic analysis [56] 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

Prototype I

Prototype II



Prototype III

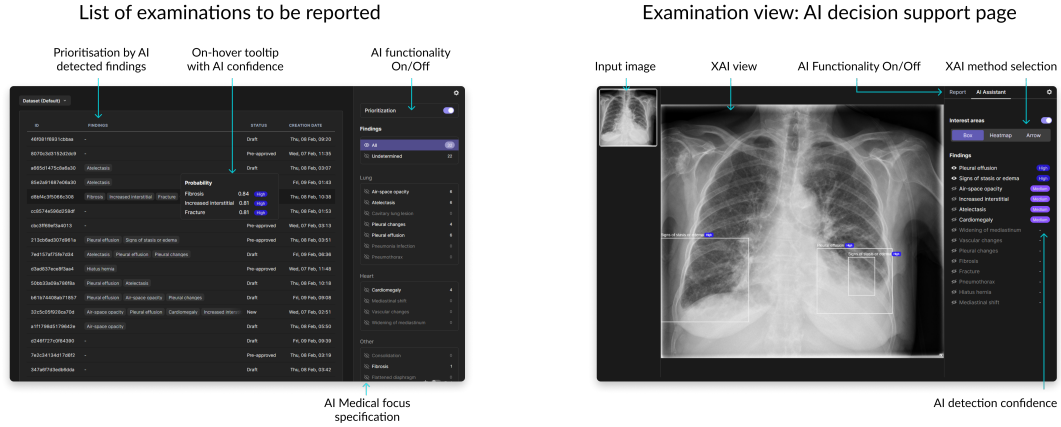


Figure 21: The collage contained three iterations of the AI prototype. Finally, four main aspects of the system can be configured: (1) AI functionality (prioritisation and decision support), (2) radiological findings detected on an X-ray, (3) the AI decision thresholds of the AI model (globally or per finding), and (4) AI explainability method.

in this study. The two authors familiarised themselves with the collected data after every iteration of the 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 [362], 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 (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 fourteen reflective themes. The fourteen themes were framed as dependencies conditioning four specific design decisions that formed an AI-based support design space.

15.4 CONFIGURING FOUR TECHNICAL DIMENSIONS OF CLINICALLY USEFUL RADIOLOGICAL AI

We conceptualised 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 Figure 22). Each of 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.

15.4.1 *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.

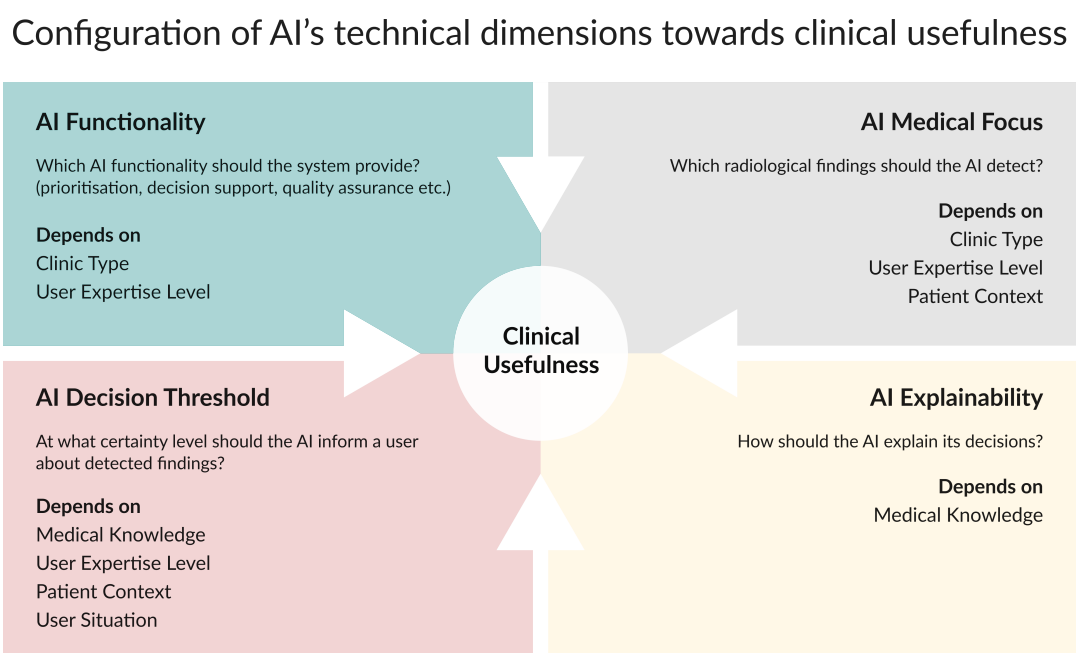


Figure 22: A matrix of technical AI dimensions that need to be configured to achieve clinical usefulness in local practice. Configuration of each of the technical dimensions is conditioned by the accompanying social dimensions of the local clinical practice. Conceptualised based on design interventions with an AI-based prototype.

CLINIC TYPE. This social dimension includes 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. Thus, it was closely related to the patient context. 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, and their health and well-being are their final goals. Thus, by extension, any system supporting healthcare professionals should support the patients. Additionally, any clinical action always depends on the patient's 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.

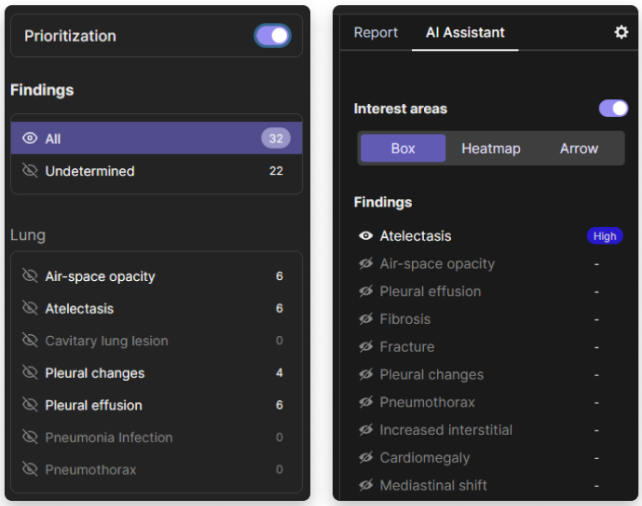


Figure 23: Two UI elements allowing to select prioritisation and decision support on their respective pages.

15.4.2 *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 Figure 23) 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 [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.* [I18, 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 imper-

ative for the creators of AI-based support systems to envision alternative motivations for the use of their systems and allow appropriate configuration.

Second, each of the healthcare centres will have different financial resources available. This factor, which has virtually never 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?* [I18, 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 5th X-ray in 20 minutes. But if I have 100 X-rays to go through, I don't want to get to the 100th 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.* [I19, 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 radiologists and 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...* [I14, 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.

Configuring AI Functionality recommendation: **Allow users to select their preferred forms of functional support provided by the AI.**

15.4.3 *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 Figure 24). 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.

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,*

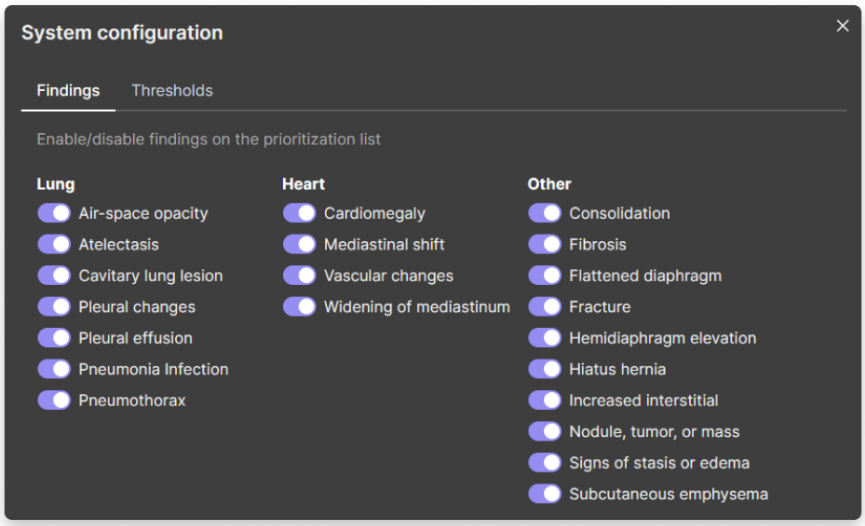


Figure 24: A configuration panel allowing users to select radiological findings detected by the AI model.

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.* [I18, 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 a finding observed in an examination of a patient who is admitted to a hospital may be expected. 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

Figure 25: 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.

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.

Configuring AI Medical Focus recommendation: **Enable users to select which radiological findings they prefer support with.**

15.4.4 *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 Figure 25). 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 consolida-

tion 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 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 people set it for themselves.* [I18, 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 pa-

tients. *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.* [I13, 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.* [I19, 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*. [I19, 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.

Configuring AI Decision Threshold recommendation: **Empower users to set personalised AI decision thresholds.**

15.4.5 *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 (Figure 26). We discovered that no single method can support the explainability of all the radiological findings.

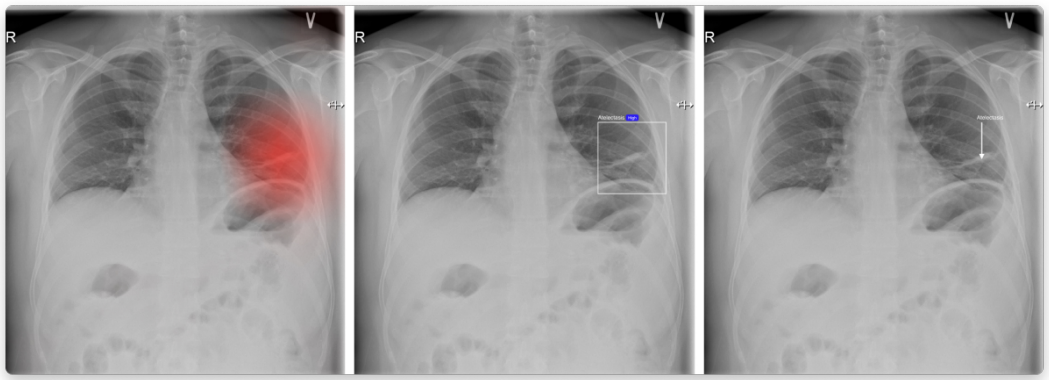


Figure 26: Different XAI methods available in the prototype. From left to right: gradient overlay, bounding box, arrow.

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...* [I13, 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.

Configuring AI Medical Focus recommendation: **Enable users to choose the most suitable XAI method for each radiological finding.**

15.5 DISCUSSION

In this paper, we investigated how to design AI for clinical usefulness in different clinical contexts of radiology practice. Based on an extended design study, we provided four practical recommendations on addressing dependencies emerging from the social dimensions of clinical practice (Figure 27). In this section, we will discuss how these recommendations may be enacted during the innovation process of clinical AI.

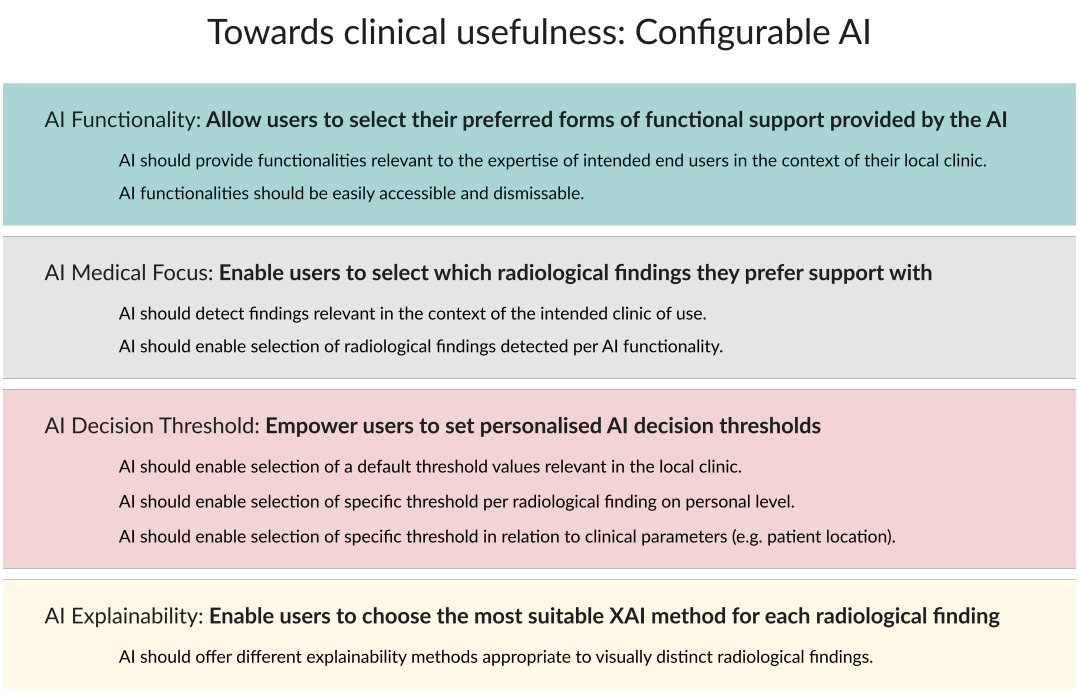


Figure 27: Four design recommendations on how to achieve clinically useful AI-based systems. Accompanied by more in-depth considerations.

15.5.1 *Allow users to select their preferred forms of functional support provided by the AI*

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 [135, 169, 263, 356]. 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 lo-

cal 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 [128, 354].

When addressing expertise-related needs for support, previous research in radiology showed that AI-based systems have different effects on junior and senior radiologists [366]. Even more, Tong et al. [328] 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 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.

The personal configuration of functionality also captures the integration - a famously difficult task when innovating clinical AI [263, 326, 357]. Many AI-based systems fail in practice due to providing support at the wrong time [28, 67, 127, 155]. Some of the AI integrations introduce a new step in the practice. A step that sometimes cannot be skipped [28]. This study shows, seconding previous research, that the integration of AI into work practices has to be flexible [52]. 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.

15.5.2 *Enable users to select which radiological findings they prefer support with*

The innovation of AI-based systems is often initiated and defined by technical opportunities, e.g., access to medical data [304, 305, 349, 365]. As such, the medical and social aspects of the systems are sometimes addressed only after the technology has gone through several rounds of development [6]. This inadvertently means that certain assumptions about the medical focus are made [361]. Present radiological AI models tend to detect findings relevant to the local radiologists involved in the data creation process [167, 247, 339]. 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 [248, 361]. 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 in contrast to systems where certain radiological findings are consistently considered urgent in patient care [348], 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.

15.5.3 *Empower users to set personalised AI decision thresholds*

Selecting AI decision threshold has significant ethical [39], performance [288], and clinical [331] 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. [191] explored how the decision threshold affects the number of false positive and false negative predictions, significantly altering a user's 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 [18, 25, 221, 325].

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 prediction. 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.

15.5.4 *Enable users to choose the most suitable XAI method for each radiological finding*

It has been long established that explainable AI fosters trust and increases the usefulness of the predictions [93, 169]. Especially in the healthcare domain, the reasoning and explanations are sometimes more valuable to end users than the predictions themselves [67, 221] or can lead to envisioning new ways of using an AI-based system altogether [172]. However, simply revealing the decision-making process of machines to humans is not enough to provide useful explanations [237]. 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.

15.6 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.

We also acknowledge the limited variability of clinical sites in Denmark compared to the visited sites in Kenya due to difficulties gaining access. 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 avenue for support to explore.

15.7 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 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.

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda." In: *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 2018-April. 2018. DOI: 10.1145/3173574.3174156.

[2] George Alexandru Adam, Chun-Hao Kingsley Chang, Benjamin Haibe-kains, and Anna Goldenberg. "Hidden Risks of Machine Learning Applied to Healthcare: Unintended Feedback Loops Between Models and Future Data Causing Model Degradation." In: *Proceedings of the 5th Machine Learning for Healthcare Conference*. Vol. 126. 2019. PMLR, 2020, pp. 710–731. URL: https://proceedings.mlr.press/v126/adam20a.html.

[3] Scott J. Adams, Robert D.E. Henderson, Xin Yi, and Paul Babyn. "Artificial Intelligence Solutions for Analysis of X-ray Images." In: *Canadian Association of Radiologists Journal* 72.1 (Feb. 2021), pp. 60–72. DOI: 10.1177/0846537120941671/ASSET/IMAGES/LARGE/10.1177{_}0846537120941671-FIG2.JPEG.

[4] Yasmeena Akhter, Richa Singh, and Mayank Vatsa. "AI-based radiodiagnosis using chest X-rays: A review." In: *Frontiers in Big Data* 6 (Apr. 2023). DOI: 10.3389/fdata.2023.1120989.

[5] Amit Alfassy, Assaf Arbelle, Oshri Halimi, Sivan Harary, Roei Herzig, Eli Schwartz, Rameswar Panda, Michele Dolfi, Christoph Auer, Kate Saenko, Peter W J Staar, Rogerio Feris, and Leonid Karlinsky. "FETA: Towards Specializing Foundational Models for Expert Task Applications." In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 29873–29888. URL: https://ai-vision-public-datasets.s3.eu.cloud-object-storage.appdomain.cloud/.

[6] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. "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*. 2019, pp. 291–300. DOI: 10.1109/ICSE-SEIP.2019.00042.

[7] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. "Power to the People: The Role of Humans in Interactive Machine Learning." In: *AI Magazine* 35.4 (Dec. 2014), pp. 105–120. DOI: 10.1609/AIMAG.V35I4.2513.

[8] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. "Guidelines for Human-AI Interaction." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019, pp. 1–13. DOI: 10.1145/3290605.3300233.

[9] Tariq Andersen, Pernille Bjørn, Finn Kensing, and Jonas Moll. "Designing for collaborative interpretation in telemonitoring: re-introducing patients as diagnostic agents." In: *International journal of medical informatics* 80.8 (Aug. 2011), pp. 112–26. DOI: 10.1016/j.ijmedinf.2010.09.010.

[10] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Enrico Coiera, and Yvonne Rogers. "Introduction to the Special Issue on Human-Centred AI in Healthcare: Challenges Appearing in the Wild." In: *ACM Transactions on Computer-Human Interaction* 30.2 (Apr. 2023), pp. 1–12. DOI: 10.1145/3589961.

[11] Ariful Islam Anik and Andrea Bunt. "Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, May 2021, pp. 1–13. DOI: 10.1145/3411764.3445736.

[12] Tomonori Aoki et al. "Clinical usefulness of a deep learning-based system as the first screening on small-bowel capsule endoscopy reading." In: *Digestive Endoscopy* 32.4 (May 2020), pp. 585–591. DOI: 10.1111/den.13517.

[13] Mohammad R. Arbabshirani, Brandon K. Fornwalt, Gino J. Mongelluzzo, Jonathan D. Suever, Brandon D. Geise, Aalpen A. Patel, and Gregory J. Moore. "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration." In: *npj Digital Medicine* 1.1 (Apr. 2018), p. 9. DOI: 10.1038/s41746-017-0015-z.

[14] 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. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." In: *Nature Medicine 2019 25:6* 25.6 (May 2019), pp. 954–961. DOI: 10.1038/s41591-019-0447-x.

[15] Lora Aroyo and Chris Welty. "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation." In: *AI Magazine* 36.1 (Mar. 2015), pp. 15–24. DOI: 10.1609/aimag.v36i1.2564.

[16] Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, Suraj Kapa, and Paul A Friedman. "An

artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction." In: *The Lancet* 394.10201 (Sept. 2019), pp. 861–867. DOI: 10.1016/S0140-6736(19)31721-0.

[17] James Auger. "Speculative design: crafting the speculation." In: *Digital Creativity* 24.1 (Mar. 2013), pp. 11–35. DOI: 10.1080/14626268.2013.767276.

[18] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. ""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, Apr. 2023, pp. 1–14. DOI: 10.1145/3544548.3581513.

[19] Ellen Balka and Ina Wagner. "Making things work: dimensions of configurability as appropriation work." In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, Nov. 2006, pp. 229–238. DOI: 10.1145/1180875.1180912.

[20] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. "Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.13 (May 2021), pp. 11405–11414. DOI: 10.1609/aaai.v35i13.17359.

[21] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance." In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), pp. 2–11. DOI: 10.1609/hcomp.v7i1.5285.

[22] Amie J. Barda, Christopher M. Horvat, and Harry Hochheiser. "A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare." In: *BMC Medical Informatics and Decision Making* 20.1 (Dec. 2020), p. 257. DOI: 10.1186/s12911-020-01276-x.

[23] Jakob E. Bardram and Claus Bossen. "A web of coordinative artifacts: Collaborative Work at a Hospital Ward." In: *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work - GROUP '05*. ACM Press, 2005, p. 168. DOI: 10.1145/1099203.1099235.

[24] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." In: *Information Fusion* 58 (June 2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[25] Sally L. Baxter, Jeremy S. Bass, and Amy M. Sitapati. "Barriers to Implementing an Artificial Intelligence Model for Unplanned Readmissions." In: *ACI Open* 04.02 (July 2020), e108–e113. DOI: 10.1055/s-0040-1716748.

[26] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. "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, Apr. 2020, pp. 1–12. DOI: 10.1145/3313831.3376718.

[27] Maura Bellio, Dominic Furniss, Neil P. Oxtoby, Sara Garbarino, Nicholas C. Firth, Annemie Ribbens, Daniel C. Alexander, and Ann Blandford. "Opportunities and Barriers for Adoption of a Decision-Support Tool for Alzheimer's Disease." In: *ACM Transactions on Computing for Healthcare* 2.4 (Oct. 2021), pp. 1–19. DOI: 10.1145/3462764.

[28] Natalie C Benda, Lala Tanmoy Das, Erika L Abramson, Katherine Blackburn, Amy Thoman, Rainu Kaushal, Yongkang Zhang, and Jessica S Ancker. ""how did you get to this number?" Stakeholder needs for implementing predictive analytics: A pre-implementation qualitative study." In: *Journal of the American Medical Informatics Association* 27.5 (May 2020), pp. 709–716. DOI: 10.1093/jamia/ocaa021.

[29] Emily M. Bender and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." In: *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), pp. 587–604. DOI: 10.1162/tacl{_}a{_}00041.

[30] Ruha Benjamin. "Race after technology: abolitionist tools for the New Jim Code." In: (2019), p. 285. URL: https://www.wiley.com/en-dk/Race+After+Technology%3A+Abolitionist+Tools+for+the+New+Jim+Code-p-9781509526437.

[31] Marc Berg. "Patient care information systems and health care work: a sociotechnical approach." In: *International Journal of Medical Informatics* 55.2 (Aug. 1999), pp. 87–101. DOI: 10.1016/S1386-5056(99)00011-8.

[32] Marc Berg, J Aarts, and J. Van der Lei. "ICT in Health Care: Sociotechnical Approaches." In: *Methods of Information in Medicine*. Vol. 42. 4. 2003, pp. 297–301. DOI: 10.1055/s-0038-1634221.

[33] Marc Berg, Chris Langenberg, Ignas v.d Berg, and Jan Kwakkernaat. "Considerations for sociotechnical design: experiences with an electronic patient record in a clinical context." In: *International Journal of Medical Informatics* 52.1-3 (Oct. 1998), pp. 243–251. DOI: 10.1016/S1386-5056(98)00143-9.

[34] Arngeir Berge, Frode Guribye, Siri-Linn Schmidt Fotland, Gro Fonnes, Ingrid H. Johansen, and Christoph Trattner. "Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls." In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. Vol. 23. ACM, July 2023, pp. 1339–1352. DOI: 10.1145/3563657.3596110.

[35] Niels van Berkel. "Making AI Work." In: *AI in Clinical Medicine*. Wiley, May 2023, pp. 448–458. DOI: 10.1002/9781119790686.ch41.

[36] Niels van Berkel, Mikael B. Skov, and Jesper Kjeldskov. "Human-AI interaction." In: *Interactions* 28.6 (Nov. 2021), pp. 67–71. DOI: 10.1145/3486941.

[37] Shlomo Berkovsky and Enrico Coiera. "Moving beyond algorithmic accuracy to improving user interaction with clinical AI." In: *PLOS Digital Health* 2.3 (Mar. 2023). Ed. by Harry Hochheiser, e0000222. DOI: 10.1371/journal.pdig.0000222.

[38] L Berlin. "Reporting the "missed" radiologic diagnosis: medicolegal and ethical considerations." In: *Radiology* 192.1 (July 1994), pp. 183–187. DOI: 10.1148/radiology.192.1.8208934.

[39] Jonathan Birch, Kathleen A. Creel, Abhinav K. Jha, and Anya Plutynski. "Clinical decisions using AI must consider patient values." In: *Nature Medicine 2022 28:2* 28.2 (Jan. 2022), pp. 229–232. DOI: 10.1038/s41591-021-01624-y.

[40] Pernille Bjørn and Nina Boulus-Rødje. "The Multiple Intersecting Sites of Design in CSCW Research." In: *Computer Supported Cooperative Work (CSCW)* 24.4 (Aug. 2015), pp. 319–351. DOI: 10.1007/s10606-015-9227-4.

[41] Pernille Bjørn and Lars Rune Christensen. "Relation work: Creating socio-technical connections in global engineering." In: *ECSCW 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work, 24-28 September 2011, Aarhus Denmark*. Springer London, 2011, pp. 133–152. DOI: 10.1007/978-0-85729-913-0{_}8.

[42] Ann Blandford. "HCI for health and wellbeing: Challenges and opportunities." In: *International Journal of Human-Computer Studies* 131 (Nov. 2019), pp. 41–51. DOI: 10.1016/j.ijhcs.2019.06.007.

[43] Ann Blandford, Jo Gibbs, Nikki Newhouse, Olga Perski, Aneesha Singh, and Elizabeth Murray. "Seven lessons for interdisciplinary research on interactive digital health interventions." In: *DIGITAL HEALTH* 4 (Jan. 2018), p. 205520761877032. DOI: 10.1177/2055207618770325.

[44] Jeanette Blomberg and Helena Karasti. "Reflections on 25 Years of Ethnography in CSCW." In: *Computer Supported Cooperative Work (CSCW)* 22.4-6 (Aug. 2013), pp. 373–423. DOI: 10.1007/s10606-012-9183-1.

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

[46] Jeanette Blomberg, Lucy Suchman, and Randall H. Trigg. "Reflections on a Work-Oriented Design Project." In: *Human–Computer Interaction* 11.3 (Sept. 1996), pp. 237–265. DOI: 10.1207/s15327051hci1103{_}3.

[47] Jeanette Blomberg, Lucy Suchman, and Randall H. Trigg. "Reflections on a work-oriented design project." In: *Hum.-Comput. Interact.* 11.3 (Sept. 1996), pp. 237–265. DOI: 10.1207/s15327051hci1103_3.

[48] T. Bodenheimer and C. Sinsky. "From Triple to Quadruple Aim: Care of the Patient Requires Care of the Provider." In: *The Annals of Family Medicine* 12.6 (Nov. 2014), pp. 573–576. DOI: 10.1370/afm.1713.

[49] Susanne Bødker, Christian Dindler, and Ole Sejer Iversen. "Tying Knots: Participatory Infrastructuring at Work." In: *Computer Supported Cooperative Work: CSCW: An International Journal* 26.1-2 (Apr. 2017), pp. 245–273. DOI: 10.1007/s10606-017-9268-y.

[50] Susanne Bødker, Pelle Ehn, Joergen Knudsen, Morten Kyng, and Kim Madsen. "Computer support for cooperative design." In: *Proceedings of the 1988 ACM conference on Computer-supported cooperative work - CSCW '88*. ACM Press, 1988, pp. 377–394. DOI: 10.1145/62266.62296.

[51] Claus Bossen and Randi Markussen. "Infrastructuring and Ordering Devices in Health Care: Medication Plans and Practices on a Hospital Ward." In: *Computer Supported Cooperative Work (CSCW)* 19.6 (Dec. 2010), pp. 615–637. DOI: 10.1007/s10606-010-9131-x.

[52] Claus Bossen and Kathleen H. Pine. "Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists." In: *ACM Transactions on Computer-Human Interaction* 30.2 (Apr. 2023), pp. 1–29. DOI: 10.1145/3569892.

[53] Bart-Jan Boverhof, W. Ken Redekop, Daniel Bos, Martijn P. A. Starmans, Judy Birch, Andrea Rockall, and Jacob J. Visser. "Radiology AI Deployment and Assessment Rubric (RADAR) to bring value-based AI into radiological practice." In: *Insights into Imaging* 15.1 (Feb. 2024), p. 34. DOI: 10.1186/s13244-023-01599-z.

[54] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out - Classification and Its Consequences*. The MIT Press, 2000. URL: https://mitpress.mit.edu/9780262522953/.

[55] Adrian Brady, Risteárd Ó Laoide, Peter McCarthy, and Ronan McDermott. "Discrepancy and error in radiology: concepts, causes and consequences." In: *The Ulster medical journal* 81.1 (Jan. 2012), pp. 3–9. URL: http://www.ncbi.nlm.nih.gov/pubmed/23536732%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3609674.

[56] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology." In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa.

[57] Meghan Brennan, Sahil Puri, Tezcan Ozrazgat-Baslanti, Zheng Feng, Matthew Ruppert, Haleh Hashemighouchani, Petar Momcilovic, Xiaolin Li, Daisy Zhe Wang, and Azra Bihorac. "Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study." In: *Surgery (United States)* 165.5 (May 2019), pp. 1035–1045. DOI: 10.1016/j.surg.2019.01.002.

[58] Nicky Britten, Rona Campbell, Catherine Pope, Jenny Donovan, Myfanwy Morgan, and Roisin Pill. "Using meta ethnography to synthesise qualitative research: A worked example." In: *Journal of Health Services Research and Policy* 7.4 (2002), pp. 209–215. DOI: 10.1258/135581902320432732.

[59] Rebekah R. Brown, Ana Deletic, and Tony H. F. Wong. "Interdisciplinarity: How to catalyse collaboration." In: *Nature* 525.7569 (Sept. 2015), pp. 315–317. DOI: 10.1038/525315a.

[60] Tom B Brown et al. "Language Models are Few-Shot Learners." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901. URL: https://commoncrawl.org/the-data/.

[61] Michael A. Bruno, Eric A. Walker, and Hani H. Abujudeh. "Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction." In: *RadioGraphics* 35.6 (Oct. 2015), pp. 1668–1676. DOI: 10.1148/rg.2015150023.

[62] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. "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, Apr. 2023, pp. 1–19. DOI: 10.1145/3544548.3581251.

[63] Monika Büscher, Satinder Gill, Preben Mogensen, and Dan Shapiro. "Landscapes of Practice: Bricolage as a Method for Situated Design." In: *Computer Supported Cooperative Work (CSCW)* 10.1 (Mar. 2001), pp. 1–28. DOI: 10.1023/A:1011293210539.

[64] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. "PadChest: A large chest x-ray image dataset with multi-label annotated reports." In: *Medical Image Analysis* 66 (Dec. 2020), p. 101797. DOI: 10.1016/j.media.2020.101797.

[65] Danilo Bzdok, Naomi Altman, and Martin Krzywinski. "Statistics versus machine learning." In: *Nature Methods* 15.4 (Apr. 2018), pp. 233–234. DOI: 10.1038/nmeth.4642.

[66] Federico Cabitza, Andrea Campagner, and Clara Balsano. "Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters." In: *Annals of Translational Medicine* 8.7 (Apr. 2020), pp. 501–501. DOI: 10.21037/atm.2020.03.63.

[67] 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. "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, May 2019, pp. 1–14. DOI: 10.1145/3290605.3300234.

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

[69] Sabrina Caldwell, Penny Sweetser, Nicholas O'Donnell, Matthew J. Knight, Matthew Aitchison, Tom Gedeon, Daniel Johnson, Margot Brereton, Marcus Gallagher, and David Conroy. "An Agile New Research Framework for Hybrid Human-AI Teaming: Trust, Transparency, and Transferability." In: *ACM Transactions on Interactive Intelligent Systems* 12.3 (Sept. 2022), pp. 1–36. DOI: 10.1145/3514257.

[70] Michelle Campbell, R Fitzpatrick, A Haines, A L Kinmonth, P Sandercock, D Spiegelhalter, and P Tyrer. "Framework for design and evaluation of complex interventions to improve health." In: *BMJ* 321.7262 (Sept. 2000), pp. 694–6. DOI: 10.1136/bmj.321.7262.694.

[71] Tara Capel and Margot Brereton. "What is Human-Centered about Human-Centered AI? A Map of the Research Landscape." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2023, pp. 1–23. DOI: 10.1145/3544548.3580959.

[72] Davide Castelvecchi. "Can we open the black box of AI?" In: *Nature* 538.7623 (Oct. 2016), pp. 20–23. DOI: 10.1038/538020a.

[73] K Charmaz. *Constructing Grounded Theory (2nd ed.)* 2014.

[74] Kathy Charmaz. *Constructing grounded theory: a practical guide through qualitative analysis*. SAGE, 2006. URL: https://cir.nii.ac.jp/crid/1130282272823478400.

[75] Haihua Chen, Jiangping Chen, and Junhua Ding. "Data evaluation and enhancement for quality improvement of machine learning." In: *IEEE Transactions on Reliability* 70.2 (June 2021), pp. 831–847. DOI: 10.1109/TR.2021.3070863.

[76] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. "How to develop machine learning models for healthcare." In: *Nature Materials* 18.5 (May 2019), pp. 410–414. DOI: 10.1038/s41563-019-0345-0.

[77] Quan Ze Chen, Tobias Schnabel, Besmira Nushi, and Saleema Amershi. "HINT: Integration Testing for AI-based features with Humans in the Loop." In: *27th International Conference on Intelligent User Interfaces*. Vol. 22. ACM, Mar. 2022, pp. 549–565. DOI: 10.1145/3490099.3511141.

[78] Yunan Chen. "Documenting transitional information in EMR." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2010, pp. 1787–1796. DOI: 10.1145/1753326.1753594.

[79] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. "Explaining Decision-Making Algorithms through UI." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019, pp. 1–12. DOI: 10.1145/3290605.3300789.

[80] Insook Cho and Insun Jin. "Responses of staff nurses to an EMR-based clinical decision support service for predicting inpatient fall risk." In: *Studies in Health Technology and Informatics*. Vol. 264. IOS Press, Aug. 2019, pp. 1650–1651. DOI: 10.3233/SHTI190579.

[81] Adele Clarke. *Situational Analysis*. SAGE Publications, Inc., July 2005. DOI: 10.4135/9781412985833.

[82] Enrico Coiera. *The last mile: Where artificial intelligence meets reality*. 2019. DOI: 10.2196/16323.

[83] Linda M Collins, Susan A Murphy, Vijay N Nair, and Victor J Strecher. "A Strategy for Optimizing and Evaluating Behavioral Interventions." In: (2005).

[84] David Coyle and Gavin Doherty. "Clinical evaluations and collaborative design: developing new technologies for mental healthcare interventions." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2009, pp. 2051–2060. DOI: 10.1145/1518701.1519013.

[85] Andrew Crabtree, Tom Rodden, Peter Tolmie, and Graham Button. "Ethnography considered harmful." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2009, pp. 879–888. DOI: 10.1145/1518701.1518835.

[86] Peter Craig, Paul Dieppe, Sally Macintyre, Susan Mitchie, Irwin Nazareth, and Mark Petticrew. *Developing and evaluating complex interventions: The new Medical Research Council guidance*. 2008. DOI: 10.1136/bmj.a1655.

[87] D Cramp and O. M. Goodyear. "Expert systems in medicine-Report on a European survey." In: *London: Healthcare Informatics Foundation* (1989).

[88] Kathrin M Cresswell, Ann Blandford, and Aziz Sheikh. *Drawing on human factors engineering to evaluate the effectiveness of health information technology*. 2017. DOI: 10.1177/0141076817712252.

[89] Karin Dembrower, Alessio Crippa, Eugenia Colón, Martin Eklund, and Fredrik Strand. "Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study." In: *The Lancet Digital Health* 0.0 (Sept. 2023). DOI: 10.1016/S2589-7500(23)00153-X.

[90] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. "Preparing a collection of radiology examinations for distribution and retrieval." In: *Journal of the American Medical Informatics Association* 23.2 (Mar. 2016), pp. 304–310. DOI: 10.1093/jamia/ocv080.

[91] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. "Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation." In: (Dec. 2021). URL: http://arxiv.org/abs/2112.04554.

[92] Advait Deshpande and Helen Sharp. "Responsible AI Systems: Who are the Stakeholders?" In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, July 2022, pp. 227–236. DOI: 10.1145/3514094.3534187.

[93] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. "Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations across the AI Lifecycle." In: *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere* 12 (June 2021), pp. 1591–1602. DOI: 10.1145/3461778.3462131.

[94] Tom Diethe, Miquel Perello Nieto, Emma Tonkin, Mike Holmes, Kacper Sokol, Niall Twomey, Meelis Kull, Hao Song, and Peter Flach. "Releasing eHealth analytics into the wild: Lessons learnt from the SPHERE project." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, July 2018, pp. 243–252. DOI: 10.1145/3219819.3219883.

[95] *Digital X-ray On-The-Go in Kenya*. URL: https://radiology.ucsf.edu/blog/digital-x-ray-go-kenya.

[96] Virginia Dignum. "Responsibility and artificial intelligence." In: *Oxford Handbook of Ethics of AI*. Ed. by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press, June 2020. Chap. 11, pp. 215–231.

[97] Steven E. Dilsizian and Eliot L. Siegel. "Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment." In: *Current Cardiology Reports* 16.1 (Jan. 2014), pp. 1–8. DOI: 10.1007/s11886-013-0441-8.

[98] Mary Dixon-Woods, Sheila Bonas, Andrew Booth, David R. Jones, Tina Miller, Alex J Sutton, Rachel L. Shaw, Jonathan A. Smith, and Bridget Young. *How can systematic reviews incorporate qualitative research? A critical perspective*. 2006. DOI: 10.1177/1468794106058867.

[99] Mary Dixon-Woods, Debbie Cavers, Shona Agarwal, Ellen Annandale, Antony Arthur, Janet Harvey, Ron Hsu, Savita Katbamna, Richard Olsen, Lucy Smith, Richard Riley, and Alex J Sutton. *Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups*. July 2006. DOI: 10.1186/1471-2288-6-35.

[100] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. "Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness." In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, June 2022, pp. 1639–1656. DOI: 10.1145/3531146.3533221.

[101] Paul Dourish. "Accounting for system behaviour: Representation, reflection and resourceful action." In: *Computers and design in context*. MIT Press Cambridge, 1997, pp. 145–170.

[102] Paul Dourish. "The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents." In: *Computer Supported Cooperative Work (CSCW)* 12.4 (Dec. 2003), pp. 465–490. DOI: 10.1023/A:1026149119426.

[103] Paul Dourish and Sara Bly. "Portholes: supporting awareness in a distributed work group." In: *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*. ACM Press, 1992, pp. 541–547. DOI: 10.1145/142750.142982.

[104] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. "UX Design Innovation: Challenges for Working with Machine Learning as a Design Material." In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017). DOI: 10.1145/3025453.

[105] Jack Dowie. "Decision Analysis: The Ethical Approach to Most Health Decision Making." In: *Principles of Health Care Ethics*. Wiley, June 2006, pp. 577–583. DOI: 10.1002/9780470510544.ch79.

[106] John J. Dudley and Per Ola Kristensson. "A Review of User Interface Design for Interactive Machine Learning." In: *ACM Transactions on Interactive Intelligent Systems* 8.2 (June 2018), pp. 1–37. DOI: 10.1145/3185517.

[107] Paul N Edwards, Steven J Jackson, Geoffrey C Bowker, and Cory P Knobel. "Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures"." In: (2007).

[108] Matthias Egger, George Davey Smith, and Keith O'Rourke. "Introduction: Rationale, Potentials, and Promise of Systematic Reviews." In: *Systematic Reviews in Health Care*. Wiley, Jan. 2001, pp. 1–19. DOI: 10.1002/9780470693926.ch1.

[109] Madeleine Clare Elish and Elizabeth Anne Watkins. *Repairing Innovation: A Study of Integrating AI in Clinical Care*. Tech. rep. Data & Society Research Institute, 2020. URL: https://datasociety.net/pubs/repairing-innovation.pdf.

[110] Arthur S Elstein and Alan Schwartz. "Clinical problem solving and diagnostic decision making: selective review of the cognitive literature." In: *BMJ (Clinical research ed.)* 324.7339 (Mar. 2002), pp. 729–32. DOI: 10.1136/bmj.324.7339.729.

[111] Yrjo Engestrom, Ritva Engestrom, and Osmo Saarelma. "Computerized medical records, production pressure and compartmentalization in the work activity of health center physicians." In: *Proceedings of the 1988 ACM conference on Computer-supported cooperative work - CSCW '88*. ACM Press, 1988, pp. 65–84. DOI: 10.1145/62266.62272.

[112] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. "Algorithmic fairness datasets: the story so far." In: *Data Mining and Knowledge Discovery 2022 36:6* 36.6 (Sept. 2022), pp. 2074–2152. DOI: 10.1007/S10618-022-00854-Z.

[113] Nicholas Fancourt et al. "Standardized Interpretation of Chest Radiographs in Cases of Pediatric Pneumonia From the PERCH Study." In: *Clinical Infectious Diseases* 64.suppl_3 (June 2017), S253–S261. DOI: 10.1093/CID/CIX082.

[114] Melanie Feinberg. "A Design Perspective on Data." In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Vol. 2017-May. ACM, May 2017, pp. 2952–2963. DOI: 10.1145/3025453.3025837.

[115] Geraldine Fitzpatrick and Gunnar Ellingsen. *A review of 25 years of CSCW research in healthcare: Contributions, challenges and future agendas*. Aug. 2013. DOI: 10.1007/s10606-012-9168-0.

[116] Alexander L Fogel and Joseph C Kvedar. "Artificial intelligence powers digital medicine." In: *npj Digital Medicine* 1.1 (2018), p. 5. DOI: 10.1038/s41746-017-0012-2.

[117] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. "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, June 2022, pp. 1362–1374. DOI: 10.1145/3531146.3533193.

[118] Diana E. Forsythe. ""It's Just a Matter of Common Sense": Ethnography as Invisible Work." In: *Computer Supported Cooperative Work (CSCW)* 8.1-2 (Mar. 1999), pp. 127–145. DOI: 10.1023/A:1008692231284.

[119] Karën Fort. *Collaborative Annotation for Reliable Natural Language Processing*. John Wiley & Sons, Inc., May 2016, pp. 1–164. DOI: 10.1002/9781119306696.

[120] Kent Fridell, Lars Edgren, Lars Lindsköld, Peter Aspelin, and Nina Lundberg. "The impact of PACS on radiologists' work practice." In: *Journal of digital imaging* 20.4 (Dec. 2007), pp. 411–21. DOI: 10.1007/s10278-006-1054-1.

[121] Vijay N. Garla and Cynthia Brandt. "Ontology-guided feature engineering for clinical text classification." In: *Journal of Biomedical Informatics* 45.5 (Oct. 2012), pp. 992–998. DOI: 10.1016/j.jbi.2012.04.010.

[122] Aimilia Gastounioti et al. "CAROTID - A web-based platform for optimal personalized management of atherosclerotic patients." In: *Computer Methods and Programs in Biomedicine* 114.2 (2014), pp. 183–193. DOI: 10.1016/j.cmpb.2014.02.006.

[123] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for datasets." In: *Communications of the ACM* 64.12 (Dec. 2021), pp. 86–92. DOI: 10.1145/3458723.

[124] Warren B. Gefter, Benjamin A. Post, and Hiroto Hatabu. "Commonly Missed Findings on Chest Radiographs." In: *Chest* 163.3 (Mar. 2023), pp. 650–661. DOI: 10.1016/j.chest.2022.10.039.

[125] R. Stuart Geiger, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. ""Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?" In: *Quantitative Science Studies* 2.3 (Nov. 2021), pp. 795–827. DOI: 10.1162/QSS{_}A{_}00144.

[126] Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S. Gene Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho. "High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks." In: (Mar. 2017). URL: http://arxiv.org/abs/1703.07047.

[127] 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. "Clinician Perception of a Machine Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock." In: *Critical care medicine* 47.11 (Nov. 2019), pp. 1477–1484. DOI: 10.1097/CCM.0000000000003803.

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

[129] Lisa Gitelman. *"Raw Data" Is an Oxymoron*. MIT Press, 2013, pp. 9–10. URL: https://nyuscholars.nyu.edu/en/publications/raw-data-is-an-oxymoron.

[130] Mark L. Graber, Nancy Franklin, and Ruthanna Gordon. "Diagnostic error in internal medicine." In: *Archives of Internal Medicine* 165.13 (July 2005), pp. 1493–1499. DOI: 10.1001/archinte.165.13.1493.

[131] Ben Green and Salomé Viljoen. "Algorithmic realism: Expanding the Boundaries of Algorithmic Thought." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Jan. 2020, pp. 19–31. DOI: 10.1145/3351095.3372840.

[132] Jonathan Grudin. "AI and HCI: Two fields divided by a common focus." In: *AI Magazine* 30.4 (Sept. 2009), pp. 48–57. DOI: 10.1609/aimag.v30i4.2271.

[133] Jonathan Grudin and Steven Poltrock. "Software engineering and the CHI & CSCW communities." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 896. Springer Verlag, 1995, pp. 93–112. DOI: 10.1007/bfb0035809.

[134] Dongxiao Gu, Changyong Liang, and Huimin Zhao. "A Case-Based Reasoning System Based on Weighted Heterogeneous Value Distance Metric for Breast Cancer Diagnosis." In: *Artif. Intell. Med.* 77.C (Mar. 2017), pp. 31–47. DOI: 10.1016/j.artmed.2017.02.003.

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

[136] Hongyan Gu et al. "Improving Workflow Integration with xPath: Design and Evaluation of a Human-AI Diagnosis System in Pathology." In: *ACM Transactions on Computer-Human Interaction* 30.2 (Apr. 2023), pp. 1–37. DOI: 10.1145/3577011.

[137] Varun Gulshan et al. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." In: *JAMA* 316.22 (Dec. 2016), p. 2402. DOI: 10.1001/jama.2016.17216.

[138] Philip J Guo, Sean Kandel, Joseph M Hellerstein, and Jeffrey Heer. "Proactive wrangling: mixed-initiative end-user programming of data transformation scripts." In: *Proceedings of the 24th annual ACM symposium on User interface software and technology* (2011).

[139] James Thomas Patrick Decourcy Hallinan, Mengling Feng, Dianwen Ng, Soon Yiew Sia, Vincent Tze Yang Tiong, Pooja Jagmohan, Andrew Makmur, and Yee Liang Thian. "Detection of Pneumothorax with Deep Learning Models: Learning From Radiologist Labels vs Natural Language Processing Model Generated Labels." In: *Academic Radiology* 29.9 (Sept. 2022), pp. 1350–1358. DOI: 10.1016/J.ACRA.2021.09.013.

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

[141] Joachim Halse and Laura Boffi. "Design Interventions as a Form of Inquiry." In: *Design Anthropological Futures*. Routledge, May 2020, pp. 89–103. DOI: 10.4324/9781003085188-8.

[142] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. "Computer Vision and Conflicting Values: Describing People with Automated Alt Text." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, July 2021, pp. 543–554. DOI: 10.1145/3461702.3462620.

[143] O. Hanseth and M. Aanestad. "Design as Bootstrapping. On the Evolution of ICT Networks in Health Care." In: *Methods of Information in Medicine* 42.04 (Feb. 2003), pp. 385–391. DOI: 10.1055/s-0038-1634234.

[144] Mark Hartswood, Rob Procter, Roger Slack, Alex Vob, Monika Buscher, Mark Rouncefield, and Philippe Rouchy. "Co-realisation: Towards a principled synthesis of ethnomethodology and participatory design." In: *Scandinavian Journal of Information Systems* 14.2 (Jan. 2002). URL: https://aisel.aisnet.org/sjis/vol14/iss2/2.

[145] Hamed Hassanzadeh, Anthony Nguyen, Sarvnaz Karimi, and Kevin Chu. "Transferability of artificial neural networks for clinical document classification across hospitals: A case study on abnormality detection from radiology reports." In: *Journal of Biomedical Informatics* 85 (Sept. 2018), pp. 68–79. DOI: 10.1016/j.jbi.2018.07.017.

[146] Joseph Hawkins. "Addressing the Shortage of Radiologists." In: *Radiology management* 23.4 (2001), pp. 26–29. URL: www.merritthawkins.com.

[147] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. "The practical implementation of artificial intelligence technologies in medicine." In: *Nature Medicine* 25.1 (Jan. 2019), pp. 30–36. DOI: 10.1038/s41591-018-0307-0.

[148] H A Heathfield and J Wyatt. "Philosophies for the design and development of clinical decision-support systems." In: *Methods of information in medicine* 32.1 (Feb. 1993), pp. 1–8. URL: http://www.ncbi.nlm.nih.gov/pubmed/8469157.

[149] Heather Heathfield. "The rise and 'fall' of expert systems in medicine." In: *Expert Systems* 16.3 (Aug. 1999), pp. 183–188. DOI: 10.1111/1468-0394.00107.

[150] Anne Henriksen and Anja Bechmann. "Building truths in AI: Making predictive algorithms doable in healthcare." In: *Information Communication and Society* 23.6 (2020), pp. 802–816. DOI: 10.1080/1369118X.2020.1751866.

[151] Morten Hertzum and Jesper Simonsen. "Configuring information systems and work practices for each other: What competences are needed locally?" In: *International Journal of Human-Computer Studies* 122 (Feb. 2019), pp. 242–255. DOI: 10.1016/j.ijhcs.2018.10.006.

[152] Charles Hill, Rachel Bellamy, Thomas Erickson, and Margaret Burnett. "Trials and tribulations of developers of intelligent systems: A field study." In: *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, Sept. 2016, pp. 162–170. DOI: 10.1109/VLHCC.2016.7739680.

[153] Birthe Hojlund Bech. "Danish Society of Radiology." In: *HealthManagement* (2006). URL: https://healthmanagement.org/c/imaging/issuearticle/danish-society-of-radiology.

[154] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards." In: (May 2018). URL: http://arxiv.org/abs/1805.03677.

[155] Judd E. Hollander, Keara L. Sease, Dina M. Sparano, Frank D. Sites, Frances S. Shofer, and William G. Baxt. "Effects of neural network feedback to physicians on admit/discharge decision for emergency department patients with chest pain." In: *Annals of Emergency Medicine* 44.3 (Sept. 2004), pp. 199–205. DOI: 10.1016/j.annemergmed.2004.02.037.

[156] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. "Shifting Concepts of Value." In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. Vol. 20. ACM, Oct. 2020, pp. 1–12. DOI: 10.1145/3419249.3420149.

[157] Naja L. Holten Møller and Signe Vikkelsø. "The Clinical Work of Secretaries: Exploring the Intersection of Administrative and Clinical Work in the Diagnosing Process." In: *From Research to Practice in the Design of Cooperative Systems: Results and Open Challenges*. Springer London, 2012, pp. 33–47. DOI: 10.1007/978-1-4471-4093-1{_}3.

[158] Andreas Holzinger. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics* 3.2 (June 2016), pp. 119–131. DOI: 10.1007/s40708-016-0042-6.

[159] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. "What do we need to build explainable AI systems for the medical domain?" In: (Dec. 2017). URL: https://arxiv.org/abs/1712.09923v1.

[160] Rory Horner. "Towards a new paradigm of global development? Beyond the limits of international development." In: *Progress in Human Geography* 44.3 (June 2020), pp. 415–436. DOI: 10.1177/0309132519836158.

[161] Eduard Hovy and Julia Lavid. "Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics." In: *INTERNATIONAL JOURNAL OF TRANSLATION* 22.1 (2010).

[162] John Hughes, Val King, Tom Rodden, and Hans Andersen. "Moving out from the control room: Ethnography in System Design." In: *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94*. ACM Press, Oct. 1994, pp. 429–439. DOI: 10.1145/192844.193065.

[163] John A Hughes and Dan Shapiro. "Faltering from Ethnography to Design." In: *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. (Dec. 1992), pp. 115–122.

[164] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. "Towards Accountability for Machine Learning Datasets." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Mar. 2021, pp. 560–575. DOI: 10.1145/3442188.3445918.

[165] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. "Towards Accountability for Machine Learning Datasets." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Mar. 2021, pp. 560–575. DOI: 10.1145/3442188.3445918.

[166] Giulia Inguaggiato, Suzanne Metselaar, Bert Molewijk, and Guy Widdershoven. "How Moral Case Deliberation Supports Good Clinical Decision Making." In: *AMA Journal of Ethics* 21.10 (Oct. 2019), pp. 913–919. DOI: 10.1001/amajethics.2019.913.

[167] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." In: *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* (Jan. 2019), pp. 590–597. URL: http://arxiv.org/abs/1901.07031.

[168] Azra Ismail and Neha Kumar. "AI in Global Health: The View from the Front Lines." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, May 2021, pp. 1–21. DOI: 10.1145/3411764.3445130.

[169] 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. "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, May 2021, pp. 1–14. DOI: 10.1145/3411764.3445385.

[170] Saahil Jain, Akshay Smit, Andrew Y. Ng, and Pranav Rajpurkar. "Effect of Radiology Report Labeler Quality on Deep Learning Models for Chest X-Ray Interpretation." In: (Apr. 2021). URL: http://arxiv.org/abs/2104.00793.

[171] Stefanie Jauk, Diether Kramer, Alexander Avian, Andrea Berghold, Werner Leodolter, and Stefan Schulz. "Technology Acceptance of a Machine Learning Algorithm Predicting Delirium in a Clinical Setting: a Mixed-Methods Study." In: *Journal of Medical Systems* 45.4 (2021). DOI: 10.1007/s10916-021-01727-6.

[172] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. "CarePre: An Intelligent Clinical Decision Assistance System." In: *ACM Trans. Comput. Healthcare* 1.1 (Mar. 2020). DOI: 10.1145/3344258.

[173] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs." In: (Jan. 2019). URL: http://arxiv.org/abs/1901.07042.

[174] Rose Johnson, Kenton O'Hara, Abigail Sellen, Claire Cousins, and Antonio Criminisi. "Exploring the potential for touchless interaction in image-guided interventional radiology." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, May 2011, pp. 3323–3332. DOI: 10.1145/1978942.1979436.

[175] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter P. Wallach. ""You Have to Piece the Puzzle Together" - Implications for Designing Decision Support in Intensive Care." In: *Proceedings of the 2020 ACM Designing Interactive*

Systems Conference. ACM, July 2020, pp. 1509–1522. DOI: 10.1145/3357236.3395436.

[176] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V. Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. "DeepMedic for Brain Tumor Segmentation." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10154 LNCS. Springer Verlag, 2016, pp. 138–149. DOI: 10.1007/978-3-319-55524-9{_}14.

[177] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. "Wrangler: Interactive Visual Specification of Data Transformation Scripts." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, May 2011, pp. 3363–3372. DOI: 10.1145/1978942.1979444.

[178] John Kang, Olivier Morin, and Julian C. Hong. "Closing the Gap Between Machine Learning and Clinical Cancer Care—First Steps Into a Larger World." In: *JAMA Oncology* 6.11 (Nov. 2020), p. 1731. DOI: 10.1001/jamaoncol.2020.4314.

[179] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. ""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, Apr. 2022, pp. 1–18. DOI: 10.1145/3491102.3517533.

[180] Naveena Karusala, Shirley Yan, Nupoor Rajkumar, Victoria G, and Richard Anderson. "Speculating with Care: Worker-centered Perspectives on Scale in a Chat-based Health Information Service." In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (Sept. 2023), pp. 1–26. DOI: 10.1145/3610210.

[181] Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. "Identifying the Intersections: User experience + research scientist collaboration in a generative machine learning interface." In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019, pp. 1–8. DOI: 10.1145/3290607.3299059.

[182] Pearse A. Keane and Eric J. Topol. "With an eye to AI and autonomous diagnosis." In: *npj Digital Medicine* 1.1 (Aug. 2018), p. 40. DOI: 10.1038/s41746-018-0048-y.

[183] Brendan S Kelly, Conor Judge, Stephanie M Bollard, Simon M Clifford, Gerard M Healy, Awsam Aziz, Prateek Mathur, Shah Islam, Kristen W Yeom, Aonghus Lawlor, and Ronan P Killeen. "Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE)." In: *European radiology* 32.11 (Nov. 2022), pp. 7998–8007. DOI: 10.1007/s00330-022-08784-6.

[184] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key challenges for delivering clinical impact with artificial intelligence." In: *BMC medicine* 17.1 (Oct. 2019), p. 195. DOI: 10.1186/s12916-019-1426-2.

[185] Finn Kensing and Joan Greenbaum. "Heritage: having a say." In: *Routledge International Handbook of Participatory Design*. Routledge, Oct. 2012, pp. 41–56. DOI: 10.4324/9780203108543-9.

[186] Finn Kensing and Andreas Munk-Madsen. "PD: structure in the toolbox." In: *Communications of the ACM* 36.6 (June 1993), pp. 78–85. DOI: 10.1145/153571.163278.

[187] Finn Kensing, Jesper Simonsen, and Keld Bodker. "MUST: A Method for Participatory Design." In: *Human–Computer Interaction* 13.2 (June 1998), pp. 167–198. DOI: 10.1207/s15327051hci1302{_}3.

[188] Finn Kensing, Jesper Simonsen, and Keld Bødker. "Participatory Design at a Radio Station 1." In: *Computer Supported Cooperative Work* 7 (1998), pp. 243–271.

[189] David Killock. "AI outperforms radiologists in mammographic screening." In: *Nature Reviews Clinical Oncology* 17.3 (Mar. 2020), pp. 134–134. DOI: 10.1038/s41571-020-0329-7.

[190] Burak Kocak, Ozlem Korkmaz Kaya, Cagri Erdim, Ece Ates Kus, and Ozgur Kilickesmez. "Artificial Intelligence in Renal Mass Characterization: A Systematic Review of Methodologic Items Related to Modeling, Performance Evaluation, Clinical Utility, and Transparency." In: *American Journal of Roentgenology* 215.5 (Nov. 2020), pp. 1113–1122. DOI: 10.2214/AJR.20.22847.

[191] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. "Will You Accept an Imperfect AI?" In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019, pp. 1–14. DOI: 10.1145/3290605.3300641.

[192] Peter Kontschieder et al. "Quantifying progression of multiple sclerosis via classification of depth videos." In: *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 17 (2014), pp. 429–437.

[193] Ross Koppel. "Role of Computerized Physician Order Entry Systems in Facilitating Medication Errors." In: *JAMA* 293.10 (Mar. 2005), p. 1197. DOI: 10.1001/jama.293.10.1197.

[194] John R. Koza, Forrest H. Bennett, David Andre, and Martin A. Keane. "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming." In: *Artificial Intelligence in Design '96*. Springer, Dordrecht, 1996, pp. 151–170. DOI: 10.1007/978-94-009-0279-4{_}9.

[195] Josua Krause, Adam Perer, and Kenney Ng. "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, May 2016, pp. 5686–5697. DOI: 10.1145/2858036.2858529.

[196] Leah Kulp and Aleksandra Sarcevic. "Design in the "medical" wild: Challenges of technology deployment." In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. Vol. 2018-April. ACM, Apr. 2018, pp. 1–6. DOI: 10.1145/3170427.3188571.

[197] Steinar Kvale. *InterViews: an introduction to qualitive research interviewing*. Sage, 1996.

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

[199] Chantelle C. Lachance and Melissa Walter. "Artificial Intelligence for Classification of Lung Nodules: A Review of Clinical Utility, Diagnostic Accuracy, Cost-Effectiveness, and Guidelines." In: *Artificial Intelligence for Classification of Lung Nodules: A Review of Clinical Utility, Diagnostic Accuracy, Cost-Effectiveness, and Guidelines* (Oct. 2020), pp. 1–23. URL: http://europepmc.org/books/NBK562929%20https://europepmc.org/article/nbk/nbk562929.

[200] Curtis P. Langlotz. "Will Artificial Intelligence Replace Radiologists?" In: *Radiology: Artificial Intelligence* 1.3 (May 2019), e190058. DOI: 10.1148/ryai.2019190058.

[201] Shaimaa Lazem, Danilo Giglitto, Makuochi Samuel Nkwo, Hafeni Mthoko, Jessica Upani, and Anicia Peters. "Challenges and Paradoxes in Decolonising HCI: A Critical Discussion." In: *Computer Supported Cooperative Work (CSCW)* 31.2 (June 2022), pp. 159–196. DOI: 10.1007/s10606-021-09398-0.

[202] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. "Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race." In: *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (July 2021), pp. 695–703. DOI: 10.1145/3461702.3462598.

[203] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. *Deep learning*. May 2015. DOI: 10.1038/nature14539.

[204] Cindy S Lee, Paul G Nagy, Sallie J Weaver, and David E Newman-Toker. *Cognitive and system factors contributing to diagnostic errors in radiology*. Aug. 2013. DOI: 10.2214/AJR.12.10375.

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

[206] Kicky G van Leeuwen, Maarten de Rooij, Steven Schalekamp, Bram van Ginneken, and Matthieu J C M Rutten. "How does artificial intelligence in radiology improve efficiency and health outcomes?" In: *Pediatric radiology* 52.11 (Oct. 2022), pp. 2087–2093. DOI: 10.1007/s00247-021-05114-8.

[207] Kicky G van Leeuwen, Steven Schalekamp, Matthieu J C M Rutten, Bram van Ginneken, and Maarten de Rooij. "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence." In: *European radiology* 31.6 (June 2021), pp. 3797–3804. DOI: 10.1007/s00330-021-07892-z.

[208] Constance D. Lehman, Robert D. Wellman, Diana S. M. Buist, Karla Kerlikowske, Anna N. A. Tosteson, and Diana L. Miglioretti. "Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection." In: *JAMA Internal Medicine* 175.11 (Nov. 2015), p. 1828. DOI: 10.1001/jamainternmed.2015.5231.

[209] Christian Leibig, Moritz Brehmer, Stefan Bunk, Danalyn Byng, Katja Pinker, and Lale Umutlu. "Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis." In: *The Lancet. Digital health* 4.7 (July 2022), e507–e519. DOI: 10.1016/S2589-7500(22)00070-X.

[210] Dana Li, Lea Marie Pehrson, Carsten Ammitzbøl Lauridsen, Lea Tøttrup, Marco Fraccaro, Desmond Elliott, Hubert D. Zajac, Sune Darkner, Jonathan Frederik Carlsen, and Michael Bachmann Nielsen. "The Added Effect of Artificial Intelligence on Physicians' Performance in Detecting Thoracic Pathologies on CT and Chest X-ray: A Systematic Review." In: *Diagnostics* 11.12 (Nov. 2021), p. 2206. DOI: 10.3390/diagnostics11122206.

[211] Dana Li et al. "Inter- and Intra-Observer Agreement When Using a Diagnostic Labeling Scheme for Annotating Findings on Chest X-rays—An Early Step in the Development of a Deep Learning-Based Decision Support System." In: *Diagnostics 2022, Vol. 12, Page 3112* 12.12 (Dec. 2022), p. 3112. DOI: 10.3390/DIAGNOSTICS12123112.

[212] Henry Lieberman, Fabio Paternò, Markus Klann, and Volker Wulf. "End-User Development: An Emerging Paradigm." In: *End User Development*. Springer Netherlands, Oct. 2006, pp. 1–8. DOI: 10.1007/1-4020-5386-X{_}1.

[213] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. "Why and why not explanations improve the intelligibility of context-aware intelligent systems." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2009, pp. 2119–2128. DOI: 10.1145/1518701.1519023.

[214] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray. "The Unified Medical Language System." In: *Yearbook of Medical Informatics* 02.01 (Aug. 1993), pp. 41–51. DOI: 10.1055/s-0038-1637976.

[215] Christopher J. Lindsell, William W. Stead, and Kevin B. Johnson. "Action-Informed Artificial Intelligence—Matching the Algorithm to the Problem." In: *JAMA* 323.21 (June 2020), p. 2141. DOI: 10.1001/jama.2020.5035.

[216] Xiaoxuan Liu et al. "Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed." In: *Nature Medicine* 25.10 (Oct. 2019), pp. 1467–1468. DOI: 10.1038/s41591-019-0603-3.

[217] Daria Loi, Christine T. Wolf, Jeanette L. Blomberg, Raphael Arar, and Margot Brereton. "Co-designing AI Futures: Integrating AI Ethics, Social Computing, and Design." In: *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. ACM, June 2019, pp. 381–384. DOI: 10.1145/3301019.3320000.

[218] P Luff, J Hindmarsh, and C Heath. *Workplace Studies: Recovering Work Practice and Informing Systems Design*. Cambridge University Press, Aug. 2000, p. 287. URL: https://kclpure.kcl.ac.uk/portal/en/publications/workplace-studies-recovering-work-practice-and-informing-systems-.

[219] Jocelyn Maclure. "The new AI spring: a deflationary view." In: *AI & SOCIETY* 35.3 (Sept. 2020), pp. 747–750. DOI: 10.1007/s00146-019-00912-z.

[220] Vincent J Major, Neil Jethani, and Yindalon Aphinyanaphongs. "Estimating real-world performance of a predictive model: a case-study in predicting mortality." In: *JAMIA Open* 3.2 (July 2020), pp. 243–251. DOI: 10.1093/jamiaopen/ooaa008.

[221] 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. "Clinician Preimplementation Perspectives of a Decision-Support Tool for the Prediction of Cardiac Arrhythmia Based on Machine Learning: Near-Live Feasibility and Qualitative Study." In: *JMIR Human Factors* 8.4 (Nov. 2021), e26964. DOI: 10.2196/26964.

[222] Nancy McCauley and Mohammad Ala. "The use of expert systems in the healthcare industry." In: *Information & Management* 22.4 (Apr. 1992), pp. 227–235. DOI: 10.1016/0378-7206(92)90025-B.

[223] Andrea McCoy and Ritankar Das. "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units." In: *BMJ Open Quality* 6.2 (Oct. 2017), e000158. DOI: 10.1136/bmjoq-2017-000158.

[224] Marry L. McHugh. "Interrater reliability: the kappa statistic." In: *Biochemia Medica* 22.3 (2012), pp. 276–282. DOI: 10.11613/BM.2012.031.

[225] Stuart McLennan, Amelia Fiske, Leo Anthony Celi, Ruth Müller, Jan Harder, Konstantin Ritt, Sami Haddadin, and Alena Buyx. "An embedded ethics approach for AI development." In: *Nature Machine Intelligence* 2.9 (July 2020), pp. 488–490. DOI: 10.1038/s42256-020-0214-1.

[226] 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. "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." In: *European radiology* 33.2 (Feb. 2023), pp. 915–924. DOI: 10.1007/s00330-022-09090-x.

[227] Teresa Mendonca, Pedro M. Ferreira, Jorge S. Marques, Andre R. S. Marcal, and Jorge Rozeira. "PH2 - A dermoscopic image database for research and benchmarking." In: *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2013, pp. 5437–5440. DOI: 10.1109/EMBC.2013.6610779.

[228] Milagros Miceli and Julian Posada. "The Data-Production Dispositif." In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (May 2022). URL: http://arxiv.org/abs/2205.11963.

[229] Milagros Miceli, Julian Posada, and Tianling Yang. "Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?" In: *Proceedings of the ACM on Human-Computer Interaction* 6.GROUP (Jan. 2022), pp. 1–14. DOI: 10.1145/3492853.

[230] Milagros Miceli, Martin Schuessler, and Tianling Yang. "Between Subjectivity and Imposition." In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (Oct. 2020), pp. 1–25. DOI: 10.1145/3415186.

[231] Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. "Documenting Data Production Processes: A Participatory Approach for Data Work." In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (Nov. 2022), pp. 1–34. DOI: 10.1145/3555623.

[232] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. "Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Mar. 2021, pp. 161–172. DOI: 10.1145/3442188.3445880.

[233] Stefania Milan and Emiliano Treré. "Big Data from the South(s): Beyond Data Universalism." In: *Television & New Media* 20.4 (May 2019), pp. 319–335. DOI: 10.1177/1527476419837739.

[234] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Jan. 2019, pp. 220–229. DOI: 10.1145/3287560.3287596.

[235] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." In: *ACM Transactions on Interactive Intelligent Systems* 11.3-4 (Dec. 2021), pp. 1–45. DOI: 10.1145/3387166.

[236] Jesper Molin, Paweł W. Woźniak, Claes Lundström, Darren Treanor, and Morten Fjeld. "Understanding Design for Automated Image Analysis in Digital Pathology." In: *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. Vol. 23-27-October-2016. ACM, Oct. 2016, pp. 1–10. DOI: 10.1145/2971485.2971561.

[237] Cecily Morrison, Kit Huckvale, Bob Corish, Richard Banks, Martin Grayson, Jonas Dorn, Abigail Sellen, and Sân Lindley. "Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis." In: *ACM Transactions on Interactive Intelligent Systems* 8.2 (June 2018), pp. 1–28. DOI: 10.1145/3181670.

[238] Urs J Muehlematter, Paola Daniore, and Kerstin N Vokinger. "Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis." In: *The Lancet Digital Health* 3.3 (Mar. 2021), e195–e203. DOI: 10.1016/S2589-7500(20)30292-2.

[239] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. "How Data Science Workers Work with Data." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019, pp. 1–15. DOI: 10.1145/3290605.3300356.

[240] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. "Designing Ground Truth and the Social Life of Labels." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, May 2021, pp. 1–16. DOI: 10.1145/3411764.3445402.

[241] Michael J. Muller and Allison Druin. "Participatory Design : The Third Space in Human–Computer Interaction." In: *The Human–Computer Interaction Handbook* (Jan. 2012), pp. 1125–1153. DOI: 10.1201/B11963-CH-49.

[242] Claudia Müller, Cornelius Neufeldt, David Randall, and Volker Wulf. "ICT-Development in Residential Care Settings: Sensitizing Design to the Life Circumstances of the Residents of a Care Home." In: *Proceedings of the SIGCHI Conference*

on Human Factors in Computing Systems. ACM, May 2012, pp. 2639–2648. DOI: 10.1145/2207676.2208655.

[243] Enid. Mumford and Mary Weir. *Computer systems in work design–the ETHICS method : effective technical and human implementation of computer systems : a work design exercise book for individuals and groups*. 1979, p. 314.

[244] 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. "Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs." In: *Radiology* 290.1 (Jan. 2019), pp. 218–228. DOI: 10.1148/RADIOL.2018180237/ASSET/IMAGES/LARGE/RADIOL.2018180237.TBL4.JPEG.

[245] Elizamary de Souza Nascimento, Iftekhar Ahmed, Edson Oliveira, Marcio Piedade Palheta, Igor Steinmacher, and Tayana Conte. "Understanding Development Process of Machine Learning Systems: Challenges and Solutions." In: *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, Sept. 2019, pp. 1–6. DOI: 10.1109/ESEM.2019.8870157.

[246] Daniel B. Neill. "Using artificial intelligence to improve hospital inpatient care." In: *IEEE Intelligent Systems* 28.2 (2013), pp. 92–95. DOI: 10.1109/MIS.2013.51.

[247] Ha Q. Nguyen et al. "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations." In: (2020). URL: https://vindr.ai/vinlab%20http://arxiv.org/abs/2012.15029.

[248] John Nolan, Peter McNair, and Jytte Brender. "Factors influencing the transferability of medical decision support systems." In: *International Journal of Bio-Medical Computing* 27.1 (Jan. 1991), pp. 7–26. DOI: 10.1016/0020-7101(91)90018-A.

[249] Kristen Nygaard and Olav Terje Bergo. *The Trade Unions-New users of researcha*. Feb. 1975. DOI: 10.1108/eb055278.

[250] Luke Oakden-Rayner. "Exploring large scale public medical image datasets." In: (July 2019). URL: http://arxiv.org/abs/1907.12720.

[251] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging." In: *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning* (Feb. 2020), pp. 151–159. DOI: 10.1145/3368555.3384468.

[252] Evan W Orenstein, Katherine Yun, Clara Warden, Michael J Westerhaus, Morgan G Mirth, Dean Karavite, Blain Mamo, Kavya Sundar, and Jeremy J Michel. "Development and dissemination of clinical decision support across institutions:

standardization and sharing of refugee health screening modules." In: *Journal of the American Medical Informatics Association* 26.12 (Dec. 2019), pp. 1515–1524. DOI: 10.1093/jamia/ocz124.

[253] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. "Realizing AI in Healthcare: Challenges Appearing in the Wild." In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, May 2021, pp. 1–5. DOI: 10.1145/3411763.3441347.

[254] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. "Rayyan—a web and mobile app for systematic reviews." In: *Systematic Reviews 2016 5:1* 5.1 (Dec. 2016), pp. 1–10. DOI: 10.1186/S13643-016-0384-4.

[255] Matthew J. Page et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." In: *Systematic Reviews* 10.1 (Dec. 2021), pp. 1–11. DOI: 10.1186/S13643-021-01626-4/FIGURES/1.

[256] Sun Young Park, Yunan Chen, and Scott Rudkin. "Technological and Organizational Adaptation of EMR Implementation in an Emergency Department." In: *ACM Transactions on Computer-Human Interaction* 22.1 (Mar. 2015), pp. 1–24. DOI: 10.1145/2656213.

[257] Sun Young Park, So Young Lee, and Yunan Chen. "The effects of EMR deployment on doctors' work practices: A qualitative study in the emergency department of a teaching hospital." In: (2011). DOI: 10.1016/j.ijmedinf.2011.12.001.

[258] Lauv Patel, Tripti Shukla, Xiuzhen Huang, David W. Ussery, and Shanzhi Wang. "Machine Learning Methods in Drug Discovery." In: *Molecules* 25.22 (Nov. 2020), p. 5277. DOI: 10.3390/molecules25225277.

[259] Vimla L. Patel, Edward H. Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R. Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. "The coming of age of artificial intelligence in medicine." In: *Artificial Intelligence in Medicine* 46.1 (May 2009), pp. 5–17. DOI: 10.1016/j.artmed.2008.07.017.

[260] Filippo Pesapane, Marina Codari, and Francesco Sardanelli. "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine." In: *European radiology experimental* 2.1 (Oct. 2018), p. 35. DOI: 10.1186/s41747-018-0061-6.

[261] Cécile Petitgand, Aude Motulsky, Jean-Louis Denis, and Catherine Régis. "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* 270 (June 2020), pp. 1001–1005. DOI: 10.3233/SHTI200312.

[262] Mårten Pettersson, Dave Randall, and Bo Helgeson. "Ambiguities, awareness and economy: a study of emergency service work." In: *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, Nov. 2002, pp. 286–295. DOI: 10.1145/587078.587118.

[263] Kathleen H. Pine and Yunan Chen. "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, Apr. 2020, pp. 1–12. DOI: 10.1145/3313831.3376818.

[264] Kathleen H. Pine and Max Liboiron. "The Politics of Measurement and Action." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Vol. 2015-April. ACM, Apr. 2015, pp. 3147–3156. DOI: 10.1145/2702123.2702298.

[265] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. "How AI Developers Overcome Communication Challenges in a Multidisciplinary Team." In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–25. DOI: 10.1145/3449205.

[266] Fernanda Polubriaginof, Nicholas P. Tatonetti, and David K. Vawdrey. "An Assessment of Family History Information Captured in an Electronic Health Record." In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2015 (2015), pp. 2035–2042. URL: /pmc/articles/PMC4765557/%20/pmc/articles/PMC4765557/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765557/.

[267] Rob Procter, Peter Tolmie, and Mark Rouncefield. "Holding AI to Account: Challenges for the Delivery of Trustworthy AI in Healthcare." In: *ACM Transactions on Computer-Human Interaction* 30.2 (Apr. 2023), pp. 1–34. DOI: 10.1145/3577009.

[268] Pranav Rajpurkar et al. "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists." In: *PLOS Medicine* 15.11 (Nov. 2018), e1002686. DOI: 10.1371/JOURNAL.PMED.1002686.

[269] Dave Randall, Richard Harper, and Mark Rouncefield. *Fieldwork for Design: Theory and Practice*. Computer Supported Cooperative Work. Springer London, 2007. DOI: 10.1007/978-1-84628-768-8.

[270] Rebecca Randell, Stephanie Wilson, and Peter Woodward. "Variations and Commonalities in Processes of Collaboration: The Need for Multi-Site Workplace Studies." In: *Computer Supported Cooperative Work (CSCW)* 20.1-2 (Apr. 2011), pp. 37–59. DOI: 10.1007/s10606-010-9127-6.

[271] Tye Rattenbury, Joe Hellerstein, Jee Rey Heer, Sean Kandel, and Connor Carreras. *Principles of data wrangling : practical techniques for data preparation*. O'Reilly Media, Inc., 2017.

[272] Michael P Recht, Marc Dewey, Keith Dreyer, Curtis Langlotz, Wiro Niessen, Barbara Prainsack, and John J Smith. "Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations." In: *European radiology* 30.6 (June 2020), pp. 3576–3584. DOI: 10.1007/s00330-020-06672-5.

[273] Madhu Reddy, Lena Mamykina, and Andrea Grimes Parker. "Designing interactive systems in healthcare." In: *Interactions* 19.1 (Jan. 2012), pp. 24–27. DOI: 10.1145/2065327.2065334.

[274] Madhu C. Reddy, Paul Dourish, and Wanda Pratt. "Temporality in Medical Work: Time also Matters." In: *Computer Supported Cooperative Work (CSCW)* 15.1 (Feb. 2006), pp. 29–53. DOI: 10.1007/s10606-005-9010-z.

[275] Sandeep Reddy, Wendy Rogers, Ville-Petteri Makinen, Enrico Coiera, Pieta Brown, Markus Wenzel, Eva Weicken, Saba Ansari, Piyush Mathur, Aaron Casey, and Blair Kelly. "Evaluation framework to guide implementation of AI systems into healthcare settings." In: *BMJ Health & Care Informatics* 28.1 (Oct. 2021), e100444. DOI: 10.1136/bmjhci-2021-100444.

[276] Dennis Reidsma and Jean Carletta. "Reliability Measurement without Limits." In: *Computational Linguistics* 34.3 (Sept. 2008), pp. 319–326. DOI: 10.1162/coli.2008.34.3.319.

[277] Mohammad Hosein Rezazade Mehrizi, Peter van Ooijen, and Milou Homan. "Applications of artificial intelligence (AI) in diagnostic radiology: a technography study." In: *European radiology* 31.4 (Apr. 2021), pp. 1805–1811. DOI: 10.1007/s00330-020-07230-9.

[278] David Ribes, Andrew S Hoffman, Steven C Slota, and Geoffrey C Bowker. "The logic of domains." en. In: *Social Studies of Science* 49.3 (June 2019). Publisher: SAGE Publications Ltd, pp. 281–309. DOI: 10.1177/0306312719849709.

[279] Abi Rimmer. "Radiologist shortage leaves patient care at risk, warns royal college." In: *BMJ* 359 (Oct. 2017), j4683. DOI: 10.1136/bmj.j4683.

[280] Michael Roberts et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans." In: *Nature Machine Intelligence* 3.3 (Mar. 2021), pp. 199–217. DOI: 10.1038/s42256-021-00307-0.

[281] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. "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, May 2021, pp. 1–14. DOI: 10.1145/3411764.3445748.

[282] P J Robinson, D Wilson, A Coral, A Murphy, and P Verow. "Variation between experienced observers in the interpretation of accident and emergency radiographs." In: *The British Journal of Radiology* 72.856 (Apr. 1999), pp. 323–330. DOI: 10.1259/bjr.72.856.10474490.

[283] Alejandro Rodriguez-Ruiz et al. "Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists." In: *JNCI: Journal of the National Cancer Institute* 111.9 (Sept. 2019), pp. 916–922. DOI: 10.1093/JNCI/DJY222.

[284] Yvonne Rogers. "Interaction design gone wild." In: *Interactions* 18.4 (July 2011), pp. 58–62. DOI: 10.1145/1978822.1978834.

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

[286] Santiago Romero-Brufau, Kirk D. Wyatt, Patricia Boyum, Mindy Mickelson, Matthew Moore, and Cheristi Cognetta-Rieke. "Implementation of Artificial Intelligence-Based Clinical Decision Support to Reduce Hospital Readmissions at a Regional Hospital." In: *Applied Clinical Informatics* 11.4 (Aug. 2020), pp. 570–577. DOI: 10.1055/s-0040-1715827.

[287] Adam Rule, Aurélien Tabard, and James D. Hollan. "Exploration and Explanation in Computational Notebooks." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Vol. 2018-April. ACM, Apr. 2018, pp. 1–12. DOI: 10.1145/3173574.3173606.

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

[289] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, May 2021, pp. 1–15. DOI: 10.1145/3411764.3445518.

[290] Sahil Sandhu, Anthony L. Lin, Nathan Brajer, Jessica Sperling, William Ratliff, Armando D. Bedoya, Suresh Balu, Cara O'Brien, and Mark P. Sendak. "Integrating a machine learning system into clinical workflows: Qualitative study." In: *Journal of Medical Internet Research* 22.11 (Nov. 2020). DOI: 10.2196/22421.

[291] Iqbal H. Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions." In: *SN Computer Science* 2.3 (May 2021), p. 160. DOI: 10.1007/s42979-021-00592-x.

[292] Shier Nee Saw and Kwan Hoong Ng. "Current challenges of implementing artificial intelligence in medical imaging." In: *Physica Medica* 100 (Aug. 2022), pp. 12–17. DOI: 10.1016/J.EJMP.2022.06.003.

[293] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. "Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development." In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–37. DOI: 10.1145/3476058.

[294] Kjeld Schmidt. "The critical role of workplace studies in CSCW." In: *Workplace Studies* (May 2000), pp. 141–149. DOI: 10.1017/CBO9780511628122.007.

[295] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. "A Step Toward More Inclusive People Annotations for Fairness." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 21. ACM, July 2021, pp. 916–925. DOI: 10.1145/3461702.3462594.

[296] Peter G. Scupelli, Yan Xiao, Susan R. Fussell, Sara Kiesler, and Mark D. Gross. "Supporting Coordination in Surgical Suites: Physical Aspects of Common Information Spaces." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2010, pp. 1777–1786. DOI: 10.1145/1753326.1753593.

[297] Ahmed Seffah, Michel C. Desmarais, and Eduard Metzker. "HCI, Usability and Software Engineering Integration: Present and Future." In: 2005, pp. 37–57. DOI: 10.1007/1-4020-4113-6{_}3.

[298] Cathrine Seidelin, Yvonne Dittrich, and Erik Grönvall. "Data Work in a Knowledge-Broker Organisation: How Cross-Organisational Data Maintenance shapes Human Data Interactions." In: *Proceedings of the 32nd International BCS Human Computer Interaction Conference, HCI 2018*. July 2018. DOI: 10.14236/ewic/HCI2018.14.

[299] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. "Fairness and Abstraction in Sociotechnical Systems." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Jan. 2019, pp. 59–68. DOI: 10.1145/3287560.3287598.

[300] Kate Sellen, Dominic Furniss, Yunan Chen, Svetlena Taneva, Aisling Ann O'Kane, and Ann Blandford. "Workshop abstract: HCI research in healthcare: Using theory from evidence to practice." In: *CHI '14 Extended Abstracts on Human Factors in*

Computing Systems. ACM, Apr. 2014, pp. 87–90. DOI: 10.1145/2559206.2559240.

[301] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. ""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*. 2020, pp. 99–109. DOI: 10.1145/3351095.3372827.

[302] Mark Sendak et al. "Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study." In: *JMIR Medical Informatics* 8.7 (July 2020), pp. 1–16. DOI: 10.2196/15182.

[303] Nigam H. Shah, Arnold Milstein, and Steven C. Bagley PhD. "Making Machine Learning Models Clinically Useful." In: *JAMA* 322.14 (Oct. 2019), p. 1351. DOI: 10.1001/jama.2019.10306.

[304] Ben Shneiderman. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." In: *International Journal of Human–Computer Interaction* 36.6 (Apr. 2020), pp. 495–504. DOI: 10.1080/10447318.2020.1741118.

[305] Ben Shneiderman. "Human-Centered Artificial Intelligence: Three Fresh Ideas." In: *AIS Transactions on Human-Computer Interaction* 12.3 (Sept. 2020), pp. 109–124. DOI: 10.17705/1thci.00131.

[306] Jesper Simonsen and Toni Robertson. *Routledge international handbook of participatory design*. 2012, pp. 1–300. DOI: 10.4324/9780203108543.

[307] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. "Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2023, pp. 1–18. DOI: 10.1145/3544548.3581075.

[308] Anders Søgaard. "Explainable natural language processing." In: *Synthesis Lectures on Human Language Technologies*. Vol. 14. 3. MORGAN & CLAYPOOL, 2021, pp. 1–123. DOI: 10.2200/S01118ED1V01Y202107HLT051.

[309] Kacper Sokol and Peter Flach. "Explainability fact sheets: A framework for systematic assessment of explainable approaches." In: *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, Jan. 2020, pp. 56–67. DOI: 10.1145/3351095.3372870.

[310] Susan Leigh Star and Anselm Strauss. "Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work." In: *Computer Supported Cooperative Work (CSCW)* 8.1-2 (Mar. 1999), pp. 9–30. DOI: 10.1023/A:1008651105359.

[311] James Stewart and Robin Williams. "The wrong trousers? beyond the design fallacy : social learning and the user." In: *User involvement in innovation processes*. Profil, 2005, pp. 39–71. URL: https://www.research.ed.ac.uk/en/publications/the-wrong-trousers-beyond-the-design-fallacy-social-learning-and-.

[312] Jonathon E. Stewart, Frank J. Rybicki, and Girish Dwivedi. "Medical Specialties Involved in Artificial Intelligence Research: Is There a Leader." In: *Tasman Medical Journal* 2.1 (Feb. 2020), pp. 20–27. URL: https://tasmanmedicaljournal.com/2020/02/medical-specialties-involved-in-artificial-intelligence-research-is-there-a-leader/.

[313] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter P C Boon, and Ellen H M Moors. "Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors." In: *European radiology* 30.10 (Oct. 2020), pp. 5525–5532. DOI: 10.1007/s00330-020-06946-y.

[314] Zhaoyuan Su. ""What is Your Envisioned Future?": Toward Human-AI Enrichment in Data Work of Asthma Care." In: *Article* 6.CSCW2 (2022). DOI: 10.1145/3555157.

[315] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. "Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. Association for Computing Machinery, Apr. 2022, pp. 1–21. DOI: 10.1145/3491102.3517537.

[316] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. "Towards A Process Model for Co-Creating AI Experiences." In: *Designing Interactive Systems Conference 2021*. ACM, June 2021, pp. 1529–1543. DOI: 10.1145/3461778.3462012.

[317] Lucille Alice. Suchman. *Human-machine reconfigurations : plans and situated actions*. Cambridge University Press, 2006, p. 314. URL: https://cds.cern.ch/record/1991545.

[318] Lucy Suchman. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, 1987. DOI: 10.11225/JCSS.6.191.

[319] Lucy Suchman, Randall Trigg, and Jeanette Blomberg. "Working artefacts: ethnomethods of the prototype." In: *The British Journal of Sociology* 53.2 (June 2002), pp. 163–179. DOI: 10.1080/00071310220133287.

[320] Lucy A. Suchman. "Office Procedure as Practical Action: Models of Work and System Design." In: *ACM Transactions on Information Systems* 1.4 (Oct. 1983), pp. 320–328. DOI: 10.1145/357442.357445.

[321] Amara Tariq, Saptarshi Purkayastha, Geetha Priya Padmanaban, Elizabeth Krupinski, Hari Trivedi, Imon Banerjee, and Judy Wawira Gichoya. "Current Clinical Applications of Artificial Intelligence in Radiology and Their Best Supporting Evidence." In: *Journal of the American College of Radiology* 17.11 (Nov. 2020), pp. 1371–1381. DOI: 10.1016/j.jacr.2020.08.018.

[322] Angelique Taylor, Michele Murakami, Soyon Kim, Ryan Chu, and Laurel D. Riek. "Hospitals of the Future: Designing Interactive Robotic Systems for Resilient Emergency Departments." In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (Nov. 2022). DOI: 10.1145/3555543.

[323] Sian Taylor-Phillips and Karoline Freeman. "Artificial intelligence to complement rather than replace radiologists in breast screening." In: *The Lancet Digital Health* 4.7 (July 2022), e478–e479. DOI: 10.1016/S2589-7500(22)00094-2.

[324] Niels K. Ternov, Anders N. Christensen, Peter J. T. Kampen, Gustav Als, Tine Vestergaard, Lars Konge, Martin Tolsgaard, Lisbet R. Hölmich, Pascale Guitera, Annette H. Chakera, and Morten R. Hannemose. "Generalizability and usefulness of artificial intelligence for skin cancer diagnostics: An algorithm validation study." In: *JEADV Clinical Practice* 1.4 (Dec. 2022), pp. 344–354. DOI: 10.1002/jvc2.59.

[325] Anja Thieme, Danielle Belgrave, and Gavin Doherty. "Machine Learning in Mental Health." In: *ACM Transactions on Computer-Human Interaction* 27.5 (Oct. 2020), pp. 1–53. DOI: 10.1145/3398069.

[326] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. "Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment." In: *ACM Transactions on Computer-Human Interaction* 30.2 (Apr. 2023), pp. 1–50. DOI: 10.1145/3564752.

[327] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. "Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making." In: *CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2022, pp. 1–17. DOI: 10.1145/3491102.3517732.

[328] Wen-Juan Tong et al. "Integration of Artificial Intelligence Decision Aids to Reduce Workload and Enhance Efficiency in Thyroid Nodule Management." In: *JAMA Network Open* 6.5 (May 2023), e2313674. DOI: 10.1001/jamanetworkopen.2023.13674.

[329] Baptiste Vasey, David A. Clifton, Gary S. Collins, Alastair K. Denniston, Livia Faes, Bart F. Geerts, Xiaoxuan Liu, Lauren Morgan, Peter Watkinson, and Peter McCulloch. "DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence." In: *Nature Medicine* 27.2 (Feb. 2021), pp. 186–187. DOI: 10.1038/s41591-021-01229-5.

[330] Himanshu Verma, Jakub Mlynar, Roger Schaer, Julien Reichenbach, Mario Jreige, John Prior, Florian Evéquoz, and Adrien Depeursinge. "Rethinking the Role of AI with Physicians in Oncology: Revealing Perspectives from Clinical and Research Workflows." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Vol. 19. ACM, Apr. 2023, pp. 1–19. DOI: 10.1145/3544548.3581506.

[331] Simon Tilma Vistisen, Tom Joseph Pollard, Steve Harris, and Simon Meyer Lauritsen. "Artificial intelligence in the clinical setting: Towards actual implementation of reliable outcome predictions." In: *European Journal of Anaesthesiology* 39.9 (Sept. 2022), pp. 729–732. DOI: 10.1097/EJA.0000000000001696.

[332] Holger Voormann and Ulrike Gut. "Agile corpus creation." In: *Corpus Linguistics and Linguistic Theory* 4.2 (Nov. 2008), pp. 235–251. DOI: 10.1515/CLLT.2008.010/MACHINEREADABLECITATION/RIS.

[333] Laura Trajber Waisbich, Supriya Roychoudhury, and Sebastian Haug. "Beyond the single story: 'Global South' polyphonies." In: *Third World Quarterly* 42.9 (Sept. 2021), pp. 2086–2095. DOI: 10.1080/01436597.2021.1948832.

[334] Dakuo Wang, Liuping Wang, and Zhan Zhang. "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, May 2021. DOI: 10.1145/3411764.3445432.

[335] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. "Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI." In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–24. DOI: 10.1145/3359313.

[336] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. "Designing theory-driven user-centric explainable AI." In: *Conference on Human Factors in Computing Systems - Proceedings*. 2019. DOI: 10.1145/3290605.3300831.

[337] Pu Wang et al. "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy." In: *Nature Biomedical Engineering* 2.10 (Oct. 2018), pp. 741–748. DOI: 10.1038/s41551-018-0301-3.

[338] Shanshan Wang, Guohua Cao, Yan Wang, Shu Liao, Qian Wang, Jun Shi, Cheng Li, and Dinggang Shen. "Review and Prospect: Artificial Intelligence in Advanced Medical Imaging." In: *Frontiers in Radiology* 1 (Dec. 2021), p. 781868. DOI: 10.3389/FRADI.2021.781868.

[339] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. "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*. 2019, pp. 369–392. DOI: 10.1007/978-3-030-13969-8{_}18.

[340] Thomas Weikert, Joshy Cyriac, Shan Yang, Ivan Nesic, Victor Parmar, and Bram Stieltjes. "A Practical Guide to Artificial Intelligence-Based Image Analysis in Radiology." In: *Investigative radiology* 55.1 (Jan. 2020), pp. 1–7. DOI: 10.1097/RLI.0000000000000600.

[341] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. "Do no harm: a roadmap for responsible machine learning for health care." In: *Nature Medicine* (2019). DOI: 10.1038/s41591-019-0548-6.

[342] Batia Mishan Wiesenfeld, Yin Aphinyanaphongs, and Oded Nov. "AI model transferability in healthcare: a sociotechnical perspective." In: *Nature Machine Intelligence 2022 4:10* 4.10 (Oct. 2022), pp. 807–809. DOI: 10.1038/s42256-022-00544-x.

[343] Terry Winograd. *Shifting viewpoints: Artificial intelligence and human-computer interaction*. Dec. 2006. DOI: 10.1016/j.artint.2006.10.011.

[344] Christine Wolf and Jeanette Blomberg. "Evaluating the Promise of Human-Algorithm Collaborations in Everyday Work Practices." In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–23. DOI: 10.1145/3359245.

[345] Christine T. Wolf. "Explainability scenarios: towards scenario-based XAI design." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Vol. Part F147615. ACM, Mar. 2019, pp. 252–257. DOI: 10.1145/3301275.3302317.

[346] World Health Organization. *World Health Statistics 2016: Monitoring Health for the Sustainable Development Goals*. World Health Organization, 2016. URL: https://books.google.com/books?id=-A4LDgAAQBAJ&newbks=1&newbks_redir=0&printsec=frontcover&pg=PP1&hl=en#v=onepage&q&f=false.

[347] Volker Wulf and Björn Golombek. "Direct activation: A concept to encourage tailoring activities." In: *Behaviour & Information Technology* 20.4 (2001), pp. 249–263. DOI: 10.1080/01449290110048016.

[348] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. "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, Apr. 2020, pp. 1–13. DOI: 10.1145/3313831.3376807.

[349] Wei Xu. "Toward human-centered AI: A Perspective from Human-Computer Interaction." In: *Interactions* 26.4 (June 2019), pp. 42–46. DOI: 10.1145/3328485.

[350] Yan Xu, Kai Hong, Junichi Tsujii, and Eric I-Chao Chang. "Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries." In: *Journal of the American Medical Informatics Association* 19.5 (2012), pp. 824–832. DOI: 10.1136/amiajnl-2011-000776.

[351] Ling Yang, Ioana Cezara Ene, Reza Arabi Belaghi, David Koff, Nina Stein, and Pasqualina Lina Santaguida. "Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review." In: *European radiology* 32.3 (Mar. 2022), pp. 1477–1495. DOI: 10.1007/s00330-021-08214-z.

[352] Qian Yang, Nikola Banovic, and John Zimmerman. "Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Vol. 2018-April. ACM, Apr. 2018, pp. 1–11. DOI: 10.1145/3173574.3173704.

[353] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. "Sketching NLP: A case study of exploring the right things to design with language intelligence." In: *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 12. 2019. ACM, 2019. DOI: 10.1145/3290605.3300415.

[354] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. "Investigating how experienced UX designers effectively work with machine learning." In: *DIS 2018 - Proceedings of the 2018 Designing Interactive Systems Conference*. 2018, pp. 585–596. DOI: 10.1145/3196709.3196730.

[355] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. "Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design." In: *Conference on Human Factors in Computing Systems - Proceedings* (2020), pp. 1–13. DOI: 10.1145/3313831.3376301.

[356] Qian Yang, Aaron Steinfeld, and John Zimmerman. "Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. ACM, May 2019, pp. 1–11. DOI: 10.1145/3290605.3300468.

[357] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. "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, May 2016, pp. 4477–4488. DOI: 10.1145/2858036.2858373.

[358] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. "Keeping Designers in the Loop." In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, July 2020, pp. 1245–1257. DOI: 10.1145/3357236.3395528.

[359] Hubert D. Zajac. "Designing ground truth for Machine Learning - conceptualisation of a collaborative design process between medical professionals and data scientists." In: *Proceedings of 20th European Conference on Computer-Supported Cooperative Work* (2022). DOI: 10.48340/ecscw2022{_}p04.

[360] Hubert D. Zając, Tariq O. Andersen, Elijah Kwasa, Ruth Wanjohi, Mary K. Onyinkwa, Edward K. Mwaniki, Samuel N. Gitau, Shawnim S. Yaseen, Jonathan F. Carlsen, Marco Fraccaro, Michael B. Nielsen, and Yunan Chen. "Towards Clinically Useful AI: Grounding AI Visions in Radiology Practices in Global South and North." Submitted February 2024 to ACM Transactions on Computer-Human Interaction.

[361] Hubert D. Zajac, Natalia R. Avlona, Finn Kensing, Tariq O. Andersen, and Irina Shklovski. "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, Aug. 2023, pp. 351–362. DOI: 10.1145/3600211.3604766.

[362] Hubert D. Zajac, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. "Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI." In: *ACM Transactions on Computer-Human Interaction* 30.2 (Apr. 2023), pp. 1–39. DOI: 10.1145/3582430.

[363] 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. "'It depends...': Configuring AI to Improve Clinical Usefulness Across Contexts." Submitted February 2024 to ACM SIGCHI Conference on Designing Interactive Systems 2024.

[364] Amy X Zhang, Michael Muller, and Dakuo Wang. "How do Data Science Workers Collaborate? Roles, Workflows, and Tools." In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020). DOI: 10.1145/3392826.

[365] 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. "Hybrid-augmented intelligence: collaboration and cognition." In: *Frontiers of Information Technology & Electronic En-*

gineering 18.2 (Feb. 2017), pp. 153–179. DOI: 10.1631/FITEE.1700053.

[366] Wenying Zhou et al. "Interpretable artificial intelligence-based app assists inexperienced radiologists in diagnosing biliary atresia from sonographic gallbladder images." In: *BMC Medicine* 22.1 (Jan. 2024), p. 29. DOI: 10.1186/s12916-024-03247-9.

[367] Xiaomu Zhou, Mark S. Ackerman, and Kai Zheng. "I just don't know why it's gone: maintaining informal information use in inpatient care." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2009, pp. 2061–2070. DOI: 10.1145/1518701.1519014.

[368] Xiaomu Zhou, Mark S. Ackerman, and Kai Zheng. "Doctors and Psychosocial Information: Records and Reuse in Inpatient Care." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2010, pp. 1767–1776. DOI: 10.1145/1753326.1753592.

[369] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. "Research through design as a method for interaction design research in HCI." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. Association for Computing Machinery, 2007, pp. 493–502. DOI: 10.1145/1240624.1240704.